Radboud University

KU LEUVEN
FACULTY OF ARTS
DEPARTMENT OF LINGUISTICS

Concept features and lexical diversity

A dialectological case study on the relationship between meaning and variation

Karlien Franco

Radboud University

KU LEUVEN
FACULTY OF ARTS
DEPARTMENT OF LINGUISTICS

Concept features and lexical diversity

A dialectological case study on the relationship between meaning and variation

Thesis presented in partial fulfillment of the requirements for the degree of Doctor in Linguistics

by Karlien Franco

Supervisors: prof. dr. Dirk Geeraerts & prof. dr. Roeland van Hout
co-supervisor: prof. dr. Dirk Speelman

2017

For Heleen, my rock.

ACKNOWLEDGEMENTS

I am indebted to my supervisors, Dirk Geeraerts, Dirk Speelman and Roeland van Hout for coming up with this research project which allowed me to spend four years exploring the Brabantic and Limburgish dialects, the many aspects of lexical diversity and carrying out all kinds of analyses (theoretically motivated or not so much). More importantly, however, I am grateful for the many inspiring conversation. Although I still cannot pretend to know how lexical variation works, what I do know, is based on what you have taught me. Many thanks to Dirk G. for letting me just show up in your office unexpectedly to ask a very vague question, while still being able to give a sensible answer that I hadn't even thought of. Thank you, Dirk S., for often surprising me during our meetings by providing a completely new perspective on what I was researching. Furthermore, for being the first to show me, as a bachelor student, the importance of sound statistical analyses for linguistics and, later, as a PhD student, for introducing me to more advanced methods. Also many thanks for allowing me to use your scripts. Thank you, Roeland, for always offering a fresh perspective and for introducing me to the broader field of dialectology and to the network of Dutch dialectologists. The conversations I had with you always made me see gaps in my analyses or open questions in the research that I had conducted. Writing up my dissertation has mostly taught me that I know very little. However, If I have learned anything, it's that I owe it all to you three.

I also want to thank the members of my supervisory committee and jury. Thank you, Benedikt Szmrecsanyi, first and foremost, for being such a supportive and encouraging colleague. I very much enjoyed collaborating with you and I also highly appreciate the support and encouragement you have offered me over the last few months. Thanks to Eveline Wandl-Vogt. I have very much enjoyed being involved in the COST action 'e-Lexicography' and you were always able to broaden my perspective and show me alternative lines of research (chapter 6, for instance, grew out of a small-scale study that I conducted in the framework of a workshop organized by you in Vienna). I am also thankful to Antal van den Bosch, for generously agreeing to be part of my jury, and to Jeroen van Craenenbroeck for agreeing to be the chair.

I am indebted to colleagues who have directly been involved in increasing the quality of this dissertation. Thank you Veronique de Tier & Jacques Van Keymeulen, for allowing me to use the data from the Dictionary of the Flemish Dialects for chapter 6 and for offering elucidating perspectives on the data. Thanks to Martijn Wieling for very helpful suggestions, mostly concerning the construction of the GAMMs in chapter 5. I am also indebted to the raters who selflessly made time to code 1825 concepts for proneness to affect: Dirk P., Dana, Igna, Isabeau and Laura.

During the past four years, I have had the pleasure to be surrounded by an inspiring group of linguists, many of whom have become friends more than colleagues. Although writing a dissertation is a solitary endeavour, I am very thankful that conducting research at QLVL and at the Department of Linguistics in general, entails being surrounded by a group of very interesting people. I would like to thank all of you and everyone else who had lunch, brunch, a picnic, a coffee break, or Friday night drinks with me for all the serious and not-so-serious conversations: Alek, Benedikt H., Dana, Dirk P., Eline, Frauke, Freek, Isabeau, Jason, Jeroen, Jocelyne, Kris, Kristina, Laura, Lena, Leonie, Melanie, Stefania, Stefano and Thomas. Special thanks to Dirk P. and Freek, for always believing in me and for inspiring me to be a much more creative person than I actually am. I'm pretty sure that half of what I've read over the last four years was recommended to me by one of you two. Thank you Eline and Kris, for the support you both offered over the last years. I very much appreciate your helpful suggestions and that you often showed me alternative ways to investigate the questions that I was interested in. Heaps of thanks to

Kristina & Leonie – the Jeeskesboomers: I will be forever grateful that the two of you decided to move to Leuven. You became not only esteemed colleagues, but also very dear friends. Your hilarious stories, never-ending support and we're-in-this-together-ness has made writing this dissertation so much easier. Special thanks also go to my fabulous office mates, Jocelyne and Isabeau (02.42!) for all the laughter, for the useful and less useful conversations, for understanding my pop-culture references (and for the face-swaps). Being surrounded by colleagues like you, makes conducting linguistic research even better.

When I was finishing this dissertation, I was constantly reminded of how supportive and understanding my friends are. You all make my life better, each in your own way. Truckloads of thanks to the 'kernraad' for understanding that "I probably won't be able to make it", but at the same time for offering me distractions when necessary: Eva, Jeroen, Laura, Laurens, Lian, Mandy, Thomas & Thomas. I have never met anyone who is more hilarious, supportive, hilarious, loving, hilarious, understanding or hilarious than you guys. I am also grateful to Lina, Nathalie & Valérie, my chickies. Thanks for understanding that I missed birthdays, housewarming parties and random get-togethers. Thank you for the coffees, girls' nights, GIFs and, most importantly, for always being there. I am also grateful for having been part of the most fun ex-leidingsploeg ever: Anne-Marie, Annelies, Katrien, Leen, Lise & Saar. I could not have imagined how supportive and encouraging you all would be and, although I've missed nearly all of our recent get-togethers, I am grateful for your understanding and support. I will go camping with you next year!

I am highly indebted to Jan, for taking care of the layout of this dissertation, especially since it was a lot of work on very short notice. Many thanks also to the rest of my family for your support, especially to oma Genk and oma Zolder. Being surrounded by such a warm family has made it much easier to finish this dissertation.

Mama, papa, thank you for being so supportive and understanding and for always listening to the important and not-so-important things when I needed to talk about them. Mama, I am especially grateful for knowing that I can call you or come over whenever, while you at the same time understand that – at least in my case – a dissertation is best written in isolation. Your endless support has made these last few months much easier. I am also grateful that you agreed to proofread my text. Papa, thanks for showing me so much support and for offering to help me in all kinds of ways. Your encouraging messages always helped me to keep going.

Finally, I am thankful to my sister, Heleen. Thank you for letting me complain about everything and nothing during the day or in the middle of the night. Thank you for letting me live on your couch when I needed to. Most importantly, thank you for always being there. You are not only the person that knows me best, but simply the best person I know.

Preface

This dissertation focuses on lexical diversity, the amount of lexical variation that a concept shows. In its simplest form, lexical diversity is defined as the number of different words or expressions that exist to refer to a particular concept, although throughout this dissertation, other, more theoretically informed operationalizations will be used as well. Importantly, the amount of lexical diversity can differ dramatically between different concepts. For instance, in the English language, more expressions exist to refer to a container for rubbish, like *garbage can*, *dustbin*, *dumpster* or *trash can*, while only one lexical item is available for a table. In this dissertation, we inquire into factors that explain such differences between concepts.

The results obtained offer a new and under-researched perspective on lexical diversity. More specifically, lexical variation is traditionally interpreted in terms of lectal features, like the geographical location of the speaker. For example, for some concepts, other variants may be available depending on the background of the language users: British speakers, for instance, would prefer the term *dustbin*, while *garbage can* is more often used by American speakers of English.[1] However, pilot studies have shown that semantic characteristics, related to the meaning of the concept for which the variants are used, influence the number of items that are available (Geeraerts & Speelman 2010 and Speelman & Geeraerts 2007, 2008). As the scope of these pilot studies was relatively limited, we aim to systematize these results by inquiring into additional sources of data. At the same time, the aim of the work presented here is broader than mere systematization, as other aspects of the relationship between meaning, in its broadest form, and lexical diversity will be investigated as well. In practice, we examine the effect of such features in dialectological varieties of Dutch, which serve as an appropriate starting point as they are characterized by a large amount of geographically stratified lexical variants.

The theoretical framework employed in this dissertation is the Cognitive Sociolinguistics paradigm, a burgeoning field of research that relies on the theoretical framework of Cognitive Linguistics, but marries it to the social dimension and ensuing variationist methodologies that are taken for granted in sociolinguistic research. This dissertation follows this paradigm by first inquiring into the effect of cognitive concept characteristics, related to the perspective on meaning that is central to Cognitive Sociolinguistics, in part 1 (chapters 3 and 4). The correlation between socio-cultural concept features and lexical diversity is examined in part 2 (chapters 5 and 6). All the analyses are, moreover, characterized by a use of solid empirical methods.

Overall, then, this dissertation contributes to the field of lexical semantics from a Cognitive Sociolinguistics perspective. It shows how, through the use of quantitative techniques on a semantically diverse dataset, aspects of the structure of lexical diversity are revealed and how an examination of dialectological data can contribute to theoretical linguistics.

1 See https://en.oxforddictionaries.com/usage/british-and-american-terms

Background

1. Introduction
A dialectological case study in Cognitive Sociolinguistics

Lexical diversity, i.e. the amount of lexical variation particular concepts show, can differ between concepts. For the concept DRUNK[1], for instance, nearly 3000 English expressions exist, including *blitzed*, *intoxicated*, *hammered* and *I'm not as think as you drunk I am* (Dickson 2009, cited in Lillo 2009). For the concept SOBER, however, a significantly smaller number of lexical items are available, like *sober* or *abstinent*. As is apparent from this example, variation in lexical diversity is influenced by the meaning of the concepts to be expressed: concepts that are prone to taboo show more variation. The finding that meaning influences lexical diversity was first confirmed on a large scale in three pilot studies (Geeraerts & Speelman 2010, Speelman & Geeraerts 2007, 2008). Importantly, however, these pilot studies not only inquired into the proneness to taboo of a particular concept, but showed that other types of meaning-related concept features (viz. features that concern the prototypical organization of the lexicon) significantly affect lexical diversity as well. Nonetheless, as these pilot studies only focused on one dialect area, viz. the Limburgish dialect area, and only took into account one universal semantic field[2], viz. the field of concepts relating to the human body, the extent to which these features are also relevant in other datasets has not yet been examined.

This dissertation is framed against the background of the Cognitive Sociolinguistics paradigm. The overarching aim is to rely on this framework to **contribute to lexical semantics** by showing that lexical diversity is not only affected by socio-cultural and situation-related properties (like a more or less formal speech situation), but also by features related to the cognitive aspects of categorization, language processing and production. In practice, then, this work aims to achieve two immediate research aims (1 and 2) and two broader goals (3 and 4):

1. to **systematize the results of the pilot studies** by establishing that the concept features that were distinguished, are stable in other semantic fields and dialect areas
2. to **elaborate on the results obtained in the pilot studies** and distinguish additional concept features that can influence the amount of lexical diversity a concept shows
3. to show how, by **combining different methods**, a better picture of the structure of lexical diversity can be obtained
4. to elucidate that introducing a **theoretical linguistic perspective to dialectological research** on lexical variation can be beneficial for both disciplines

In this first chapter, we will first provide an overview of the research paradigm that was initiated in Geeraerts, Grondelaers & Bakema (1994), which examines the relationship between a Cognitive Linguistic perspective on meaning and the structure of lexical variation. In 1.2, the social turn in Cognitive Linguistics will be discussed. This section concludes with an overview of the Cognitive Sociolinguistics paradigm, which forms the framework of this dissertation. In 1.3 some of the factors that have been distinguished in dialectological research concerning the way dialectal variants spread across geographical space, and into reasons why differences between dialects occur, are outlined. This will indicate that a more theoretically-informed approach to lexical diversity in dialect data is necessary. The final section

1 In this dissertation, SMALL CAPS are used to indicate concepts, while *italics* are used for the lexical item used to refer to these concepts.

2 Throughout this dissertation we follow the distinction proposed by Lyons (1977: 266-269), between lexical fields, which only include simplex lexical items, and semantic fields, which are broader as other types of constructions, like complex expressions, are included as well. Consequently, we use the expression 'semantic field' to emphasize the fact that the concepts included in such fields are semantically related, although no formal boundaries are imposed on the lexical items that are considered.

(1.4) of this chapter elaborates in more detail on the aims that are central to this study and provides an overview of the four case studies presented in this dissertation. Central to these case studies are two different types of concept characteristics that influence lexical diversity. Part 1 (chapters 3 and 4) predominantly focuses on cognitive concept features, related to the organization of the lexicon, and in part 2 (chapters 5 and 6) socio-cultural concept features, which concern the experience and environment of the dialect speaker, take up a central position.

1.1 THE COGNITIVE LINGUISTIC APPROACH TO MEANING AND LEXICAL VARIATION

This dissertation fits into the Cognitive Sociolinguistics paradigm, and, more specifically, into the Cognitive Linguistic approach to lexical variation that was first initiated in Geeraerts et al. (1994). The following paragraphs elaborate in more detail on the core aspects of Cognitive Linguistics that are relevant for this dissertation. In 1.1.1, we will provide a brief overview of the main characteristics of the research into lexical variation as initiated in Geeraerts et al. (1994) and show how this dissertation contributes to this paradigm. In the following paragraphs, the perspective on meaning that underlies this line of research is described: a maximalist, usage-based view (1.1.2) and, related to this, a prototype-theoretical view on categorization (1.1.3).

However, this introductory chapter is selective, as it only briefly introduces Cognitive Linguistics and only focuses on the aspects that are relevant for the framework of this dissertation. For more extensive, exhaustive and detailed introductions into Cognitive Linguistics, the reader is referred to Croft & Cruse (2004), Dancygier (2017), Dirven & Verspoor (2004), Dubrowska & Divjak (2015), Evans & Green (2006), Geeraerts (2006) and Geeraerts & Cuyckens (2007).

1.1.1 Meaning and variation in Geeraerts, Grondelaers & Bakema (1994)

Geeraerts et al. (1994), which can be considered a study in Cognitive Sociolinguistics *avant la lettre*, examine the structure of lexical variation in the use of clothing terminology in Dutch. Crucially, it is the first study to systematically emphasize the importance of two distinctions. On the one hand, it shows that in order to obtain a full picture of the structure of lexical variation, semasiological research should be complemented with an onomasiological approach. The semasiological perspective examines the range of applications of a particular expression. Semasiology is, for this reason, often defined as research into the **meaning** of a particular item: given a particular word or expression, what are the referents to which the word applies? In the case of the word *monitor*, for instance, a semasiological analysis would reveal that it can refer both to a YOUTH LEADER, and to a COMPUTER SCREEN. The onomasiological perspective, however, investigates **naming** rather than meaning. An onomasiological approach, thus, starts from a particular (type of) referent or concept and determines which names exist or can be used to refer to the referent. For instance, an onomasiological analysis of the concept DRUNK would reveal that a large set of words can be used for this concept, including *blitzed*, *intoxicated* and *hammered*.

On the other hand, this study was the first to make the importance of the interaction between four different types of lexical variation for the structure of the lexicon explicit (also see Geeraerts 2016). First, it examines **semasiological variation**, the situation where a single lexical item can refer to more than one referent. For example, the lexical item *pants* can both be used to refer to a TWO-LEGGED TYPE OF OUTER GARMENT (IN GENERAL), but also to a more specific referent, viz. MEN'S UNDERWEAR. The second and third types of lexical variation that are distinguished concern two varieties of onomasiological variation: conceptual onomasiological variation and formal onomasiological variation. **Conceptual onomasiological variation** concerns the situation where "a referent or type of referent may be named by means of various conceptually distinct lexical categories" (Geeraerts et al. 1994: 3-4). For example, to refer to a pair of BLUE JEANS, a language user can either choose to select a lexical item belonging to the concept BLUE JEANS and use a word like *jeans* or *blue jeans*, or (s)he can conceptualize the referent as a type of PANTS, a superordinate concept, and call the denotatum *trousers* or *pants*. **Formal onomasiological variation** occurs when a choice has to be made between different synonymous expressions for a referent. In the blue jeans example, this would involve determining the relative frequency of the terms *jeans* versus *blue jeans* versus *trousers* versus *pants*. Finally, it shows how **contextual variation** can be at play both at the semasiological and onomasiological level. Contextual variation (also called **speaker and situation related variation**) is broadly defined: it includes both the relatively stable lectal properties of the interlocutors involved (like their gender or their nationality), but also transient situation-related features, like the register of the speech event (Geeraerts, Kristiansen & Peirsman 2010: 8). For the (onomasiological) BLUE JEANS example, for instance, contextual variation may take the form of determining whether older people are more likely to refer to the concept as *blue jeans*.

Additionally, from the outset, the research programme has always emphasized the importance of using solid empirical data, most frequently in the form of large corpora, for

the analysis of linguistic phenomena. Furthermore, it is characterized by a strong focus on the use of innovative quantitative techniques.

The research paradigm was subsequently extended in several ways. Geeraerts, Grondelaers & Speelman (1999) provide a study of the diachronic lexical convergence and divergence between the two standard varieties of Dutch, Belgian and Netherlandic Dutch, and of the internal stratification of Belgian Dutch in terms of Colloquial Belgian Dutch. In this study, the notion of an **onomasiological profile** was introduced (it was further developed in Speelman, Grondelaers & Geeraerts 2003). The onomasiological profile of a concept can be considered as a way to comply with the notion of a linguistic variable and the principle of accountability in sociolinguistics (Labov 1969: 737), by taking into account *all* the different synonyms that can be used to refer to the same concept.[3] Furthermore, the relative frequency of each variant is included in the calculation of the onomasiological profile to determine the degree to which the lexical items take up a strong position *vis-à-vis* alternatives for the concept. This allows for a quantification of the degree of homogeneity, or standardization, in the use of lexical variants for a particular concept. Other studies that follow the research paradigm discussed above and that have taken into account the stratification of the base dialects, which are the main focus of this dissertation, include Grieve, Speelman & Geeraerts (2011) and Szelid & Geeraerts (2008).

Three pilot studies situated in this research paradigm have further examined the relationship between meaning and variation in dialect data (Geeraerts & Speelman 2010, Speelman & Geeraerts 2007, 2008). These studies show that properties of the meanings (i.e. concepts) to be expressed influence the amount of lexical diversity that is found in the semantic field 'the human body' in the Dictionary of the Limburgish Dialects. However, these pilot studies only took into account one semantic field (viz. the human body) and one dialect area (viz. the Limburgish dialects of Dutch). As a result, the extent to which these features are relevant and stable in other semantic fields is unknown. The immediate research goals of this dissertation are, therefore, to show that the results of these pilot studies are stable in other semantic fields and to elaborate on the results that were obtained.

Crucially, the view on meaning employed in the studies mentioned above and in this dissertation, relies on the perspective on meaning of the Cognitive Linguistics paradigm. However, in Geeraerts et al. (1994) and in later studies, the main aim was to describe the systematicity in lexical variation in terms of the interaction between semantic and lectal variables. This dissertation contributes to the research programme by examining a different aspect of the structure of lexical variation. More specifically, we focus on how the interaction between semantic and lectal variables also influences the *amount* of lexical variation that is found.

1.1.2 A maximalist approach to meaning

Following from "The cognitive commitment", which entails that the description of human language should be congruent with what is known about cognition within and outside of linguistics (Lakoff 1990), Cognitive Linguistics aims to be a psychologically plausible model. Language is, thus, primarily studied as a means to communicate – a way to convey and process meaning. Furthermore, a maximalist, non-reductionist perspective on linguistic knowledge is assumed. Language systems are considered to be "reflections of general conceptual organization, categorization principles, processing mechanisms, and experiential and environmental influences" (Geeraerts & Cuyckens 2007: 3).

The movement, therefore, places a large emphasis on meaning. Four characteristics of the view on meaning in Cognitive Linguistics are often mentioned: meaning is dynamic and flexible, perspectival, encyclopaedic and non-autonomous, and based on usage and experience (Geeraerts 2006).

First, the dynamism and flexibility of meaning shows up in the recognition that linguistic meaning cannot be described in terms of necessary-and-sufficient conditions (see below). Furthermore, through the flexibility of meaning, general patterns in semantic change can be distinguished (e.g. Geeraerts 1997: chapter 3, Sweetser 1990).

Second, meaning is considered to be perspectival because, through general cognitive abilities, like perception and attention, interactions with the world are not categorized in an objective way, but they are subjectively construed. In the work of Leonard Talmy, for instance, patterns of conceptual organization expressed by grammatical constructions across languages are distinguished (e.g. Talmy 1978, 2006 [1988]). He shows, among other examples, that many languages contain the option of specifying 'plexity', the quantity of equivalent elements, by means of grammatical morphemes. In its simplest form, i.e. for objects that are matter, this coincides with the linguistic category of 'number' which is expressed in English with morphemes ∅ and -s for singular and plural. In contrast, languages generally do not contain the option to grammatically indicate the colour of an object, although this can be expressed by means of lexical items. In the lexicon, the perspectival nature of meaning is reflected by the fact that alternative expressions

3 An onomasiological profile is, at the same time, more general than a linguistic variable, because other types of (lexical) variation can be taken into as well. It can, for instance, also be used for conceptual onomasiological variation (e.g. the same clothing item is first referred to as a *crop top* and later with a term for a more general concept, like *T-shirt*).

for a single concept can focus on different referential properties of the concept. Names for different plants, for instance, can highlight various aspects of the plants like their colour, e.g. *white clover*, or their shape, e.g. *curly dock*. Overall, this indicates that linguistic expressions reflect the fact that the structure of meaning depends on the subjective interpretation and construal of the referential world.

Third, from the non-autonomous nature of linguistic knowledge, it follows that meaning is encyclopaedic. Meaning does not exist in isolation from other cognitive capacities, but it is also related to the "physical, social and linguistic context of speech events" (Langacker 1988b: 6). Such an encyclopaedic perspective on meaning has been advocated in research on category formation and embodiment theory. Research along these lines has shown that biological properties of a human being influence the way he structures his environment and uses language. Johnson (1987: XV), for instance, writes that metaphors like MORE IS UP, are based on natural everyday bodily experiences: "If you add more liquid to a container, the level goes up. If you add more objects to a pile, the level goes up. MORE and UP are correlated in our experience in a way that provides a physical basis for our abstract understanding of quantity." Additionally, as communication involves social and often culturally-dependent interactions between people, another aspect of the encyclopaedic nature of meaning concerns socio-cultural knowledge. While this was explicitly included in the theory of embodiment from the outset (Lakoff & Johnson 1980: 117-119), according to Rohrer (2007), it has sometimes remained underspecified or even neglected (also see Zlatev 1997). Sinha & Jensen de López (2001: 20), for instance, argue that the embodiment thesis "does not […] specify in which ways these two aspects [i.e. biological properties and socio-cultural environment, KF] of the organism's environment relate to each other; nor in what respects varying social environments may give rise to varying experiences; nor the extent to which such varying experiences may be relevant to the categories which are formed as a (partial) consequence of such experience." In response to this lacuna, they demonstrate that repeated exposure to culturally-bound practices induces cognitive differences that are further entrenched by, and reflected in, diverging patterns in language use (ibid., Jensen de López, Hayashi & Sinha 2005).

The final central tenet of Cognitive Linguistics is that language is based on experience with actual linguistic events and that through experience, linguistic units become more entrenched (Langacker 1988a). A dialectic relationship is, thus, presumed between language use and the grammatical system: grammar is shaped by usage. In this sense, language is emergent, as linguistic knowledge is prone to constant reorganization under the influence of experience with language use (Schmid 2016b). For this reason, many scholars in Cognitive Linguistics have argued that to understand how language is constructed and how it varies or changes, naturalistic language data have to be taken into account (e.g. Bybee & Hopper 2001, Dabrowska & Divjak 2015, Geeraerts 2005, Kemmer & Barlow 2000, Tomasello 2001). This premise naturally holds for the study of linguistic meaning, which forms the core of the focus of Cognitive Linguistics.

1.1.3 A prototype-theoretical perspective on conceptual organization

One particular aspect of the flexible nature of meaning requires more elaboration against the background of this dissertation. More specifically, in the Cognitive Linguistics movement, a prototype-theoretical view on language is taken for granted. This view first originated in psychological research on categorization (see Rosch 1978, 1987 [1974]), which showed that many categories, i.e. sets of objects that are considered to be equivalent, cannot realistically be described in terms of necessary-and-sufficient conditions. Instead, category membership is characterized by gradedness concerning the degree to which particular referents are typical for the category, and by indeterminacy concerning category boundaries. Fuzziness at the boundaries of a category shows up in examples like 'rice' for the category of 'vegetables' (is rice a type of vegetable?) and 'necklace' for the category of 'clothing' (is a necklace a piece of clothing?). Differences in typicality within a category are reflected by the fact that people consistently agree on the degree of representativeness of exemplars of categories and the same exemplars are chosen as typical by different subjects (Mervis & Rosch 1981). Cars are, for instance, considered more typical types of vehicles than elevators across subjects. Rosch & Mervis (1975) show that the degree of typicality of an item is, furthermore, negatively correlated with the number of attributes (like 'you eat it' for the categories of vegetables or fruits) shared with other members belonging to the category. Furthermore, such differences also affect other measures like reaction times (subjects respond more quickly in verification experiments of category membership to more representative exemplars) and learning and development (for instance, category membership is established first for more representative exemplars).

The findings of Eleanor Rosch and colleagues were taken up in linguistics (Geeraerts 2010: 187-192): as mentioned above, the flexibility of meaning is nowadays considered a basic property of conceptual organization in Cognitive Linguistics. Four types of prototype effects have been distinguished that can be organized along two dimensions, although they need not always co-occur. First, prototypicality effects relate to non-equality (i.e. differences in typicality or representativeness between referents belonging

	extensional (exemplars)	intensional (definition)
non-equality (salience, core/periphery)	differences of typicality and membership salience	clustering into family resemblances
non-discreteness (demarcation, flexibility)	fuzziness at the edges, membership uncertainty	absence of necessary and sufficient conditions

TABLE 1.1
Cross-classification of prototypicality effects (Geeraerts 2010: 189)

to a category), on the one hand, and non-discreteness (i.e. indeterminacy concerning the boundaries of a category) on the other. Second, prototypicality effects work both on the intensional (i.e. definitional) and extensional (i.e. referential) level. These two dimensions cross-classify (Table 1.1, from Geeraerts 2010: 189).

A standard example concerns the category of fruits. First, extensional non-equality shows up in the fact that an apple is a more prototypical type of fruit than a pineapple. Extensional non-discreteness is exemplified by considering an olive: should, in folk classifications, olives, which are the edible seed-bearing parts of the olive tree, like apples are the edible seed-bearing parts of apple trees, be considered as types of fruit? Third, intensional non-equality shows up in the fact that more typical types of fruit share more common attributes typical for the category, while less prototypical members also share attributes with other categories. For example, a lemon, a less typical type of fruit, is not sweet, whereas more typical fruits (apples, bananas, strawberries) are. Additionally, prototypical categories are structured in the form of family resemblances with attributes shared by some, but not all members. Finally, intensional non-discreteness shows up in the fact that no set of necessary and sufficient conditions can be constituted for the category 'fruit'. For instance, assuming that the most typical types of fruits are characterized by the fact that they are generally cooked by adding sugar, also allows for other types of objects, like rhubarb, to be included in the category (see Wierzbicka 1985, in Geeraerts 2010: 135-137, which contains a more comprehensive and precise overview of the absence of necessary and sufficient conditions for this category). Imposing a more strict definitional bound on the category, like 'fruit is characterized by the presence of a skin that is harder than their soft inside', excludes objects that, in a folk model of the category, are considered types of fruit, such as strawberries.

The prototype-theoretical model was subsequently extended in linguistics to account for polysemy effects and language change. For example, a radial categories model was developed by Claudia Brugman (an overview can be found in Brugman & Lakoff 1988) and popularized through Lakoff (1987). This model was developed as a way to describe the structure of different senses within a category. It can account for the fact that, although language change is unpredictable, it is principled, systematic and recurrent.

Another extension of prototype theory in linguistics, which is crucial to this dissertation, is situated at the level of the distinction between semasiology and onomasiology. Traditionally, research in (extensions of) prototype theory takes a semasiological perspective: it concerns the range of applications of a particular expression. Crucially, in Geeraerts et al. (1994), it was shown that prototypicality effects are also at play at the onomasiological level. Non-equality shows up in the fact that some concepts are more onomasiologically salient than others: the concept COFFEE, for instance, is more salient, i.e. psychologically more entrenched, than the concept BUBBLE TEA (a type of beverage that generally consists of a mixture of tea and milk, often with tapioca and other flavours added). Non-discreteness can show up in two ways (ibid.: 122): in the form of demarcation problems among semantic fields (e.g. where does the semantic field of vegetables end and that of fruits begin?) or in the form of fuzziness at the edges of concepts belonging to the same semantic field (e.g. in the semantic field of weather phenomena, where does the concept TO RAIN HEAVILY end and the concept TO STORM begin?).

1.2 THE SOCIAL TURN IN COGNITIVE LINGUISTICS

Various researchers have argued that Cognitive Linguistics should incorporate lectal stratification, be it along a social, pragmatic, cultural or other axis (e.g. Croft 2009, Dabrowksa 2015, Geeraerts et al. 1994, Harder 2003, Majid & Burenhult 2014, Schmid 2016b). Geeraerts (2005), for instance, shows that the need for this social or, more broadly, lectal dimension involves two aspects. First the usage-based nature of Cognitive Linguistics entails that the movement should take into account lectal variation for two reasons. On the one hand, any type of usage data is lectally stratified along some dimension (dialectal, sociolectal, ideolectal etc.). On

the other hand, lectal differences reflect differences in meaning, in the sense that, for instance, *hammered* is predominantly restricted to informal contexts, whereas *intoxicated* carries a more formal connotation. Second, the social nature of language in turn implies that empirical methods should be used as language users cannot necessarily, on the basis of introspection, recall contextual variation due to the diasystemic nature of language.

In this paragraph, a brief overview of two approaches that aim to introduce a social perspective into Cognitive Linguistics, is provided. In 1.2.1 the socio-pragmatic view on entrenchment and conventionalization described by Schmid is presented. This model offers an interesting approach to linguistic variation, because the redefinition of these two concepts, entrenchment and conventionalization, will be relevant for the remainder of this dissertation. However, this section will also show that the model as presented by Schmid is inadequate, because an onomasiological perspective needs to be included. This perspective is present in the Cognitive Sociolinguistics paradigm (1.2.2), which forms the framework of this dissertation.

1.2.1 A socio-pragmatic model of entrenchment and conventionalization

Schmid (2015) provides an account of the interaction between the socio-pragmatic functions of language on the one hand, and cognitive processes, on the other. More specifically, he describes a model in which the psychological entrenchment of linguistic knowledge, a cognitive process, is separated from the socio-pragmatic process of conventionalization in a speech community, although they are both related to language usage. Entrenchment, a concept introduced by Langacker (1987), involves the routinization and holistic processing of a construction in the mind of a single language user.[4] Conventionalization (also see Langacker 1987:65-66, 2008:21) concerns the coordination of linguistic knowledge and practices within a speech community through language use: it relates to a (tacit) norm of collective agreement. Both entrenchment and conventionalization are effects of, and processes that influence, repeated exposure to language use. As a result, the emergent nature of language is central to the model. Crucially, Schmid's model also explicitly takes into account socio-pragmatic functions of language by including social forces, like setting, participants and intentions, and pragmatic forces, like social network, solidarity and prestige, which determine the way entrenchment and conventionalization processes interact with language use and how they, ultimately, shape language.

While the model proposed by Schmid offers an interesting perspective on language variation and change, in the framework of this dissertation, some further elaboration needs to be considered (for discussion, see Geeraerts 2016 and Schmid 2016a: 447). More specifically, taking a usage-based perspective on linguistic research entails that the stratification of linguistic variants should be taken into account explicitly (see above). Furthermore, because this dissertation investigates variation in lexical diversity, an onomasiological perspective to entrenchment and conventionalization is necessary: to quantify the amount of lexical diversity in a particular speech community, *all* the lexical items that exist to refer to a particular concept, together with their lectal stratification, need to be taken into account. In Schmid's model, such a perspective is implicitly included because he argues that different social and pragmatic forces have an effect on the entrenchment and conventionalization processes that shape language. If such forces influence the way "alternative ways of saying the same thing" (Labov 1969: 738) are produced in the relevant and specific (socio-pragmatic) environment (e.g. two participants of a particular age and gender involved in an informal face-to-face conversation), an onomasiological perspective underlies the model. However, as the main focus of the model is to describe the degree of entrenchment and conventionalization of a single linguistic variant, regardless of other alternative variants that may exist, it is ultimately a semasiological one.

Crucially, the notions of entrenchment and conventionalization, as described in Langacker's and Schmid's models, do interact with the amount of lexical diversity a concept shows from an onomasiological perspective: in a particular socio-pragmatic environment, which is in this dissertation predominantly equated with the specific location of a dialect speaker, a certain linguistic variant will probably be the preferred one to refer to a concept, both in cognitive and social terms. Consequently, throughout this dissertation, this terminology will be used to refer to the degree of social (conventionalization) and psychological (entrenchment) embeddedness of a particular linguistic variant from a semasiological perspective. However, to investigate variation in the amount of lexical diversity on a higher level, i.e. in an entire dialect area, a broader perspective is necessary.

4 In Schmid's (2015: 10) view, the notion of entrenchment is somewhat broader as it encompasses association, routinization and schematization.

1.2.2 Dialectological lexical diversity from the Cognitive Sociolinguistics perspective

Such a broader, onomasiological perspective has been put forward in the framework of Cognitive Sociolinguistics (see the volumes edited by Kristiansen & Dirven 2008, Geeraerts, Kristiansen & Peirsman 2010 and Pütz, Robinson & Reif 2014). Cognitive Sociolinguistics combines the theoretical framework developed in Cognitive Linguistics with the tradition of variationist sociolinguistics of employing solid empirical methods to examine the socio-cultural position of a language user as a correlate of language variation and change. Thus, aside from (1) relying on the theoretical framework elaborated in Cognitive Linguistics in general, research in this prototypically structured paradigm is characterized by two other features (Kristiansen & Dirven 2008: 5-6). More specifically, research in Cognitive Sociolinguistics (2) explicitly includes the social dimensions of variation; and (3) uses solid empirical methods. Consequently, as a sociolinguistically informed, usage-based perspective is taken, an onomasiological orientation is preferred, as this can be considered as a generalization of the notion of the linguistic variable as it is defined in sociolinguistics (Geeraerts et al. 2010).

Although few studies in Cognitive Sociolinguistics have focused on variation in dialectal varieties (an exception is Szelid & Geeraerts 2008), the combination of the theoretical framework employed in the paradigm, the attention for lectal variation and the use of empirical methods offers a solid way to research this type of variation.

First, through the combination of a maximalist view on meaning and a focus on social aspects of variation, attention can be paid to the nature of lexical diversity. More specifically, it can be argued that different operationalizations of diversity entail a different perspective. First, lexical diversity can be measured by merely taking into account how many different expressions exist to convey a particular meaning (possibly by also including the number of available tokens). In the context of this dissertation, this entails counting the number of unique types that occur per concept. However, such a measure does not take into account the way in which the distribution of particular lexical variants reflects social structure. For instance, Figure 1.1 shows the spatial distribution of the lexical items *x* and *y* for two concepts A and B in a fictional dialect area. For both concepts only two lexical items exist. However, as the figure shows, this does not necessarily entail that the geographical spread of these items shows the same degree of homogeneity. More specifically, each item used for concept A is limited to a specific and well-defined geographical region (e.g. the northern part of Belgium and the Netherlands). Lexical item *x* for concept B occurs almost throughout the entire region, while lexical item *y* only has a very limited geographical scope for this concept (it only occurs near the border with a neighbouring language, e.g. German). The figure indicates that the amount of lexical diversity that concept B shows is smaller than the lexical diversity of A, because for B, a much smaller geographical region is delineated from the rest of the dialect area, resulting in a higher degree of homogeneity overall. Consequently, lexical diversity also concerns the degree of homogeneity in the distribution of a particular variant within a particular region, i.e. the extent to which one or a few lexical variants are the preferred form over their alternatives. As mentioned above, in the line of research developed in Geeraerts et al. (1999), a profile-based approach has been employed to model such an onomasiological perspective in corpus data.

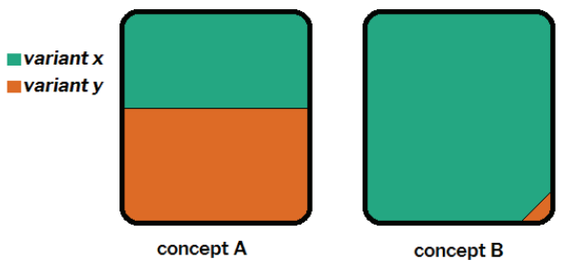


FIGURE 1.1
Differences in the homogeneity of the distribution of two lexical variants for two concepts in a fictitious dialect area

Additionally, researching the structure of the interaction between the "varieties of variation" (Geeraerts et al. 1994: 1), like formal, conceptual and contextual onomasiological variation, is central to the Cognitive Sociolinguistics paradigm (Geeraerts et al. 2010). In this dissertation, we focus on the interaction between formal onomasiological variation, on the one hand, and contextual variation, on the other. More specifically, we examine variation in the amount of lexical diversity for a set of concepts but do not take into account the way these referents are conceptualized. The speaker-related perspective takes up a central position because the geographical location of a speaker is central to the type of variation that we find. However, speaker-related features will only explicitly be included in the analyses in part 2 of this dissertation.

Finally, this dissertation also complies with the final characteristic of Cognitive Sociolinguistics, the use of solid empirical methodologies, by relying on a large dataset of naturalistic dialect data to investigate variation in the amount of lexical diversity a concept shows. More specifically, we analyse questionnaire data available in the digitized databases of the Dictionaries of the Brabantic and Limburgish Dialects (see chapter 2). Additionally, to establish the validity of the results that we obtain, we examine a sufficiently large

number of thematically diverse semantic fields. Finally, the analyses in this dissertation rely on the use of appropriate inferential statistical methods to establish, with a reasonable amount of certainty, that concept-related features influence variability in the amount of lexical diversity.

1.3 APPROACHES TO LEXICAL DIVERSITY IN TRADITIONAL DIALECTOLOGY

Dialectology[5] is, obviously, the field of linguistic research that concerns the study of dialects. Although often a broad perspective on the concept of a dialect is taken (e.g. Chambers & Trudgill 1980: 3), in this dissertation we use a more limited interpretation. We consider what is often referred to as the 'base dialects' of Dutch, which are prototypically characterized by geographical stratification. Bloomfield (1958 [1933]: 325), for instance, argues that "[e]very village, or, at most, every cluster of two or three villages, has its local peculiarities of speech." In practice, for the dialects headed by the Dutch standard languages, we take into account the predominantly spoken set of varieties that are highly spatially and, to some extent, sociolinguistically bound and that have been losing ground in favour of more regionally dispersed varieties of Dutch, like *Tussentaal* 'intermediate language', *Polderdutch* and Standard Dutch in the Netherlands (Geeraerts & Van de Velde 2013). Furthermore, this dissertation investigates a historical set of language varieties, as it studies these dialects in the form that they were spoken at the beginning of the 20th century. For instance, we will be concerned with the local dialects spoken in those days in locations like Zonhoven, Sittard and Margraten in the Limburgish dialect area and with the dialects of Brecht, Mechelen and Almkerk in the Brabantic region.

Research into the geographical stratification of dialects gained ground in the wake of the neogrammarian search for exceptionless sound laws (Bloomfield 1954 [1933]: 321-345, Chambers & Trudgill 1980: 18-23). This gave way to an interest in the systematicity with which linguistic variants spread across geographical space, which resulted in the construction of a large number of dialect surveys and atlases, like the German surveys of Georg Wenker, which were later edited and published by Ferdinand Wrede, and the French dialect atlas project edited by Jules Giliéron. However, relatively quickly, such dialectological enterprises showed that the spatial distribution of variants is highly heterogeneous. Although some general processes have been noticed, the search for complete regularity in the spread of sound laws was rapidly abandoned. For example, Kloeke (1927 in Bloomfield 1954 [1933]: §19.4) was the first to notice an example of the process of lexical diffusion ("every word has its own history"; ibid.: 328) in the pronunciation of the vowels in *muis* 'mouse' and *huis* 'house' in Dutch. The dialectal pronunciation of these vowels is geographically stratified: in the eastern periphery, the traditional Germanic pronunciation with [u:] is retained in both words, while in other locations, [y:] is used for both (in a third set of locations, in the centre of the language area, the present-day standard Dutch diphthong [œy:] occurs in both words, but this is a later development). However, additionally, in three relatively large regions towards the east of the language area, the pronunciation of both vowels differs: [mu:s, hy:s]. It is argued that the use of [y:] spread from cultural centres in the west of the language area, first Flanders and later Holland, from the Middle Ages onwards. Towards the east of the language area, another group, the North Germanic Hanse, took up a culturally significant position, and they used the Germanic variant [u:]. Additionally, [y:] was considered the more elegant variant, which was even hypercorrectly used in words that do not generally have [y:] like [vy:t] for [vu:t] 'foot'. Crucially, the larger geographical expansion of the [y:] variant in the word *house* has been explained as a semantic frequency-related phenomenon: *house* probably occurs more, especially in formal speech, while the use of the word *mouse* is probably more limited to homely situations. Consequently, following from the confrontation with these types of non-exceptionless phonological developments, dialect geography became a field of research in its own right (Nerbonne & Kretzschmar 2003): it became the aim of the dialectologist to distinguish smaller dialect areas, characterized by a certain degree of linguistic uniformity, within a larger heterogeneous region.

The traditional method to distinguish dialect areas from each other relies on maps of the variants used in particular locations for a set of linguistic variables, like concepts or the pronunciation of particular words. These maps are then interpreted by the dialectologist, in order to identify possible isoglosses, borders for a single linguistic variable that distinguish areas where a particular variant is used from regions with a different variant (examples of such dialect maps in Dutch dialectology are, for example, presented in Blancquaert & Pée 1925-1982). If on a large number of maps, the same isoglosses can be found (i.e. bundles of isoglosses), these can often be interpreted as dialect borders. Recently, advances in dialectometry have automatized and objectified this process by taking into account quantitative measures of the linguistic distance between different locations (examples for the Dutch language area are Heeringa 2014, Hoppenbrouwers & Hoppenbrouwers 2001 and Wieling 2012).

5 In line with the definition of a dialect outlined in this paragraph, we use this term in a strict way, viz. referring to the study of the geographically stratified base dialects.

Alternatively, dialect areas have also been distinguished on the basis of subjective, perceptual distances of dialect speakers (Preston 1999, Weijnen 1946).

While the main aim of dialect geography has, thus, become to distinguish different varieties within a broader area, some dialectologists have also offered explanations for the dialect areas that they find. Such explanations generally inquire into the ease of spatial diffusion of particular variants and take the form of an interaction between geographical distance and social or political factors, like mobility, population size, different types of language learning or the presence of a language border (Britain 2002, 2011, Chambers & Trudgill 196-204, Labov 2007, Weinreich, Labov & Herzog 1968: 153-155). Additionally, according to Weinreich (1968: 1-2), borrowing between dialects of the same language is influenced by the same features as borrowing from genetically unrelated languages, although the potential of interference is smaller between related languages as they share fewer mutually exclusive forms.

Occasionally, other types of influential factors are mentioned as well. For instance, in his seminal overview of Dutch dialectology, Weijnen (1966: 70-149) lists a number of other factors that explain the limited diffusion of linguistic variants and the presence of particular isoglosses in the local Dutch dialects. Such factors include, but are not limited to, topological features, historical boundaries between cultural communities and folkloristic practices. For example, he provides a list of examples where the presence of a swamp, dune, hill or bay coincides with borders between dialects (also see Nichols 2013). For instance, in the province of Zeeland in the Netherlands the names for a PURSE differ per island: *borre* is used on the island of Goeree, *bozze* on Schouwen and Southern Beveland and *beuze* on Walcheren and Noord-Beverland. Furthermore, dialect borders sometimes coincide with folkloristic borders. According to Weijnen, this is the case for the distinction between the East and West Brabantic dialect areas: not only are these dialects linguistically differentiated, they are also characterized by different folkloristic practices, which are often limited to one of the two dialect regions.

Additionally, some semantic factors that influence lexical differences between dialects have been described as well. First, differences between the structure of local communities or between local practices may cause variation in the amount of diversity within a particular *set* of related concepts (also see Chambers & Trudgill 1980: 120-123). According to Weijnen (1967: 337), some dialects of fisherman communities of Dutch have a separate name for the eldest brother, whereas most other dialects do not make this distinction, because this brother needs to take up a caring position if the father drowns while fishing (a few other examples are mentioned in Goossens 1964 and in Weijnen 1967: 337). Furthermore, fear of homonymy and polysemy have been put forward as mechanisms of lexical change, under the assumption that one form should exist to express one meaning (the principle of isomorphy). The traditional example with regard to avoidance of homonymy concerns the words for ROOSTER in the Gascon dialect area, described by Gilliéron (Gilliéron & Roques 1912 in Geeraerts 2010: 62-63). Regular sound change caused the Latin words for ROOSTER, *gallus*, and CAT, *cattus*, to merge in these dialects as *gat*. As this type of homonymy is problematic in an agrarian society, the word for ROOSTER was replaced by *azan* (the local variant of *faisan* 'pheasant'), *bigey* (probably related to *vicaire* 'curate') or *poule* (<Latin *pullus*) in the Gascon dialects. Crucially, the use of two of these words (*azan* and *bigey*) directly coincides with the isogloss delineating the area where *gallus* and *cattus* would have merged due to the regular sound change (although the word *pullus* also occurs in some locations where the sound change would not have taken place). Fear of polysemy has, for instance, been described for the Dutch dialects in Goossens (1972: 94-96). He argues that in geographical transitional areas that are characterized by the fact that, on each side of the transitional area, the same word form is used to refer to different concepts, the language users avoid using this, from their point of view, polysemous word form and rely on other lexical items to express the two meanings. For instance, in the Westbrabantic and Limburgish dialects, lexical items related to *lopen* 'to run', like *lopig*, are used to refer to an ANIMAL IN HEAT (see Table 1.2 for a schematic representation, based on Geeraerts 1986: 165). In the east of this dialect region, *lopig* only refers to cows in heat, while in the west, this item is solely used for dogs. In the transitional zone between the eastern and the western area, lexical items related to *lopen* are not used in any of these meanings. Instead, the dialect speakers from the transitional area rely on other words, like *heet* 'hot' or *willig* 'wanting', to avoid the polysemous *lopen*-related word forms.[6]

Crucially, the fact that none of the factors listed above truly inquire into variation in lexical diversity may be related to several characteristics of traditional research in dialectology. First, the features that have been distinguished are fragmentary, as each explanation only accounts for the distribution of one or of a small set of linguistic variables. Research in dialectology has not yet been able to distinguish features that may correlate with lexical diversity in the lexicon as a whole, because a systematic study on an extensive

6 A refinement of the principle of avoidance of polysemy is presented in Geeraerts (1986, 1987, 1997: chapter 4), which takes into account the flexible and polysemous structure of the lexicon.

	western area	transitional area	eastern area
IN HEAT (OF COWS)	*lopig*	*willig*	*willig*
IN HEAT (OF DOGS)	*vuil*	*heet*	*lopig*

TABLE 1.2
Schematic representation of lexical items used for concepts related to ANIMAL IN HEAT *(based on Geeraerts 1986: 165)*

number of semantic fields has not yet been conducted (although the recent study by Pickl 2013 can be considered an exception). Second, some explanations that are provided above only concern the spatial distribution of a single variant to refer to a linguistic variable, but quantifications of the frequency of alternative expressions are missing. These two problems can probably be explained by the fact that traditional research in dialectology does not always directly take into account insights into the structure of language obtained in Cognitive Linguistics (but see Swanenberg 2000 and later for an exception).

1.4 AIMS AND OUTLINE

The central aim of this dissertation is to **contribute to the knowledge of the structure of lexical variation in the framework of Cognitive Sociolinguistics**. More specifically, the overarching goal is to show that lexical diversity is not only affected by socio-cultural differences between the speakers of a dialect, but also by features related to the cognitive aspects of categorization, language processing and production. While all our case studies rely on dialectological data, we believe that they offer a first, but clear indication of the importance of the interaction between these features on the structure of the lexicon.

On the basis of this overarching aim, two immediate research goals can be distinguished. First, we aim to **systematize the results obtained in the pilot studies** discussed above (Geeraerts & Speelman 2010, Speelman & Geeraerts 2007, 2008). More specifically, the concept features that were distinguished in the pilot studies, viz. lack of onomasiological salience (non-equality), onomasiological vagueness (non-discreteness) and sensitivity to negative affect, were shown to correlate positively with lexical diversity in the semantic field of the human body in the Dictionary of the Limburgish Dialects. By including a diverse set of other semantic fields in the analysis and data from a different dialect area, we examine whether they also have a significant effect in other semantic fields and if the effect is the same. This forms the topic of chapter 3.

The second immediate research goal is to **elaborate on these results** by distinguishing other concept-related features that influence lexical diversity as well. In practice, we elaborate on the results from the pilot studies in four ways. In part one, we mostly focus on cognitive concept features, related to the organization of the lexicon. First, taking into account a larger dataset offers the possibility of determining the extent to which the relative impact of the concept features differs per semantic field. By selecting semantic fields that differ with regard to their average degree of concreteness and universality, we can explain inter-field differences by relying on these factors. These kinds of differences are examined in chapters 3 and 4. Second, we aim to determine whether the concept features have the same effect on different aspects of lexical diversity. More specifically, in chapter 4, we conduct an exploratory analysis of whether the concept features used in chapter 3 have the same effect on geographical homogeneity in the spread of lexical variants as they do on the number of heteronymous expressions that are available.

Whereas in part 1, the concept features are considered to be stable across dialect speakers, in part 2, this view is abandoned and attention is paid to the socio-cultural and historical background of the speech community and of the language user. In chapter 5, we examine to what extent the geographical stratification of the lexical variants in the database can be explained by social and semantic features. More specifically, we focus on the spatial distribution of lexical borrowings from different source languages and from different semantic fields. The final case study, presented in chapter 6, also takes into account socio-cultural differences, but it takes another perspective. It relies on referential, objectively collected extra-linguistic data, to gauge the effect of concept frequency in the environment of a dialect speaker on lexical diversity. More specifically, we zoom in on the correlation between geographically determined variation in the natural occurrence of plants in the northern part of Belgium and the amount of lexical diversity each plant shows. Overall, then, these four case studies reveal the interplay between lectal and semantic features.

Furthermore, this study aims to meet two broader research goals. Methodologically, a variety of inferential statistical techniques are used in the analyses, which take into account the geographical stratification of the data in different ways. Additionally, throughout the case studies, we use and compare different operationalizations of lexical diversity. In chapter 3, diversity consists of a composite variable that includes both the number of unique types per concept

and the way these variants are scattered across space in a less homogeneous way. In chapter 4, these two aspects are explicitly taken apart. This reveals that not every cognitive concept feature influences lexical diversity in the same way. In chapter 5, lexical diversity is operationalized by taking into account whether a concept is expressed with a native or non-native variant. The dependent variable then takes the form of the geographical distribution of native *vis-à-vis* non-native material. The results of the analyses indicate that lexical borrowing is heavily influenced by socio-cultural history, but that the meaning of the concept to be expressed plays a large role as well. Chapter 6 ties the other chapters together in several ways. First, it explicitly compares several operationalizations of lexical diversity, which were problematized in chapters 3 and 4. It relates to chapter 5 as it also includes properties of the socio-cultural environment of a dialect speaker. However, while such features are in chapter 5 used in the interpretation of the results, in chapter 6 they are included as independent variables in the analysis. Consequently, we aim to show that by **combining the results obtained by using a variety of techniques, a better picture of lexical variation can be acquired**.

Finally, the second broader research goal is to show how **a dialectological case study can contribute to theoretical linguistics and vice versa**. On the one hand, we provide a systematic study of lexical diversity in dialect data, which has not yet been done before. In this sense, we contribute to the field of dialectology. Although traditional dialect-geographical analyses of the Brabantic and Limburgish dialect areas will only play a very minor role in this dissertation, chapter 5 will also provide an aggregate geographical picture of the stratification of loanwords *vis-à-vis* native dialectal variants. On the other hand, through our examination of historical dialectal data, we reveal aspects of the structure of the lexicon that may be relevant for differently stratified varieties as well. More specifically, our study contributes to the research programme initiated in Geeraerts et al. (1994) as it aims to show that the interaction between semantic and lectal features also influences variation in the *amount* of lexical variation a concept shows.

Before explaining the case studies in greater detail, chapter 2 provides an overview of the data used in this dissertation. In chapter 7, a discussion of the results obtained in the four pilot studies follows. This discussion will examine to what extent the aims outlined in this introductory chapter are met by the case studies and suggest directions for future research.

2. Data

The data used in this dissertation come from the digitized databases of the Woordenboek van de Brabantse Dialecten '*Dictionary of the Brabantic Dialects*' (WBD) and from the Woordenboek van de Limburgse Dialecten '*Dictionary of the Limburgish Dialects*' (WLD). This chapter will provide a concise overview of these data. Section 2.1 gives a brief overview of the history of these large-scale regional dictionary projects. In 2.2, the lexicon included in the dictionaries is discussed. In 2.3, the structure of the digitized databases will be described. Section 2.4 outlines the advantages and limitations of using these data.

2.1 A BRIEF HISTORY OF THE (SOUTHERN) DUTCH DICTIONARY PROJECTS

This section is to a large extent based on the description of the Southern Dutch dictionary projects in Kruijsen (1996), Kruijsen & Van Keymeulen (1997), Van Keymeulen (1992: 1-3), the preliminary introduction to the WBD (Weijnen & Van Bakel 1967) and the introduction to the WLD (Weijnen, Goossens & Goossens 1983). The recent history of the dictionary projects is described in De Vriend & Swanenberg (2006) and in De Vriend et al. (2006).

Professor A. Weijnen, the founding father of the dictionaries of the Brabantic and Limburgish dialects and an expert in dialectology, first conceived of a plan to compose a Brabantic dialect dictionary when he became professor of Dutch and Indo-Germanic linguistics at the University of Nijmegen in 1958. In the first half of the 1960s, owing to the fact that the national Dutch council for scientific research (ZWO, now NWO) subsidised the dictionary project, Jan van Bakel became co-editor. This resulted in the establishment of a research centre, the *Nijmeegse Centrale voor Dialect- en Naamkunde* 'Centre for dialectology and onomastics of Nijmegen' (NCDN). This centre started the distribution of large-scale written questionnaires to elicit lexical variation in the Brabantic dialect area. The first volume of the WBD appeared in 1967, together with a general introduction to the dictionary (Weijnen & Van Bakel 1967). The final volumes were published in 2005.

The way Weijnen envisaged the dictionary was innovative in two respects. First, he attached great importance to the scientific quality and the systematicity of the project. In Weijnen (1975b: 87 [1967]), for instance, he writes that the importance of the Dictionary of the Brabantic Dialects lies in the fact that:

> *"het een thesaurus wil zijn, een schat van alle woorden die wij in het brabants (zowel van benoorden als bezuiden de rijksgrens) sinds de eerste gepubliceerde brabantse dialectwoordenlijst […] hebben kunnen achterhalen."*
> *(it aims to be a thesaurus, a treasure-house of all the words that we have been able to record in the Brabantic dialect (north and south of the state border) since the publication of the first Brabantic dialect lexicon.)*

Second, he vehemently argued that the dialect data should be presented in an onomasiological (or 'ideological', see for instance Weijnen 1975a, [1961]) way and that a semantic organisation based on the way concepts are related to each other in human life should be used, rather than the traditional alphabetical order.

The dictionary consists of three parts, agrarian terminology, specialized non-agrarian terminology and general vocabulary. For each part, a number of thematic volumes are published, like farming tools, butcher & bakery or the human body. Furthermore, each thematic volume is organized onomasiologically: for each concept, a list of heteronyms (i.e. geographically stratified synonymous lexical items) is provided.

At the beginning of the 1960s, work on the Limburgish dictionary started as well at the NCDN. The concrete initiative to construct a Limburgish dictionary came from

Weijnen as well, in collaboration with J. L. Pauwels in Leuven and with support from W. Roukens[1]. The structure of the Limburgish dictionary has been identical to the Brabantic one from the outset. The questionnaires were also distributed by the NCDN, supported by the *Zuidnederlandse Dialectcentrale* 'Southern Dutch dialect centre' in Leuven for the distribution of the questionnaires in the Belgian province of Limburg. The first volume of the WLD appeared in 1983, together with the general introduction to the dictionary (Weijnen, Goossens & Goossens 1983). The last volume was finished in 2008.

Two other dictionary projects were created following the example of the WBD and WLD. The first volume of the *Woordenboek van de Achterhoekse en Liemerse dialecten* 'Dictionary of the Dialects of Liemers and Achterhoek' (WALD) was published in 1984. Work on the *Woordenboek van de Vlaamse Dialecten* 'Dictionary of the Flemish[2] Dialects' was initiated by W. Pee at the University of Ghent in 1972. As a result, the dialect lexicon of the entire area below the 'big rivers', i.e. the southern part of the Dutch language area which comprises the south of the Netherlands and the northern part of Belgium, is recorded.

In 1989, the collaboration between the editors of the WBD, WLD and WVD was formalized in a coordinating project to achieve a common policy concerning the publication of the data (Nederlandse Taalunie 1990). This has resulted in a congruent publication of the dialect data of the three dictionaries for the third part of the dictionaries, which comprises the general, non-specialist vocabulary. Furthermore, due to technological innovations, most of the dialect data has by now been digitized (Belemans & Goossens 2000: 11-12 and Kruijsen 2001: VI-IX). From the publication of the general vocabulary onwards, the details of the paper version of each volume of the WBD and WLD were made available in (online) databases, which are referred to as 'materiaalbases'. In these databases, a dictionary user can, for instance, look up information about the exact geographical distribution of a specific lexical item whereas before, only the distribution of concepts was available in the dictionary.[3] An offline version of these databases forms the data that are used in this dissertation, although separate datasets are used for the Brabantic and Limburgish data. One of the recent collaborative aims of the three southern Dutch dictionary projects is to create a common, integrated database.[4]

2.2 WHAT KIND OF DIALECT LEXICON?

In the introduction to the WLD, the editors state that the goal of the dictionary is to make an inventory, as exhaustive as possible, of the Limburgish lexicon in every situation in which the base dialects function as the means of communication (Weijnen, Goossens & Goossens 1983: 4). Neither the introduction to the WLD, nor to the WBD is explicit about what the prototypical dialect speaker looks like (e.g. when and where did he grow up? what is his profession? what is his age and gender?), although this is relevant from a sociolinguistic or variationist perspective. According to Kruijsen (1996), the traditional dialect lexicon in the dictionaries is perceived as the early 20th century common lexical norm of a large part of the speech community, which is still known by the oldest generation in the second half of the 20th century. As a result, the dialect dictionaries not only serve a linguistic aim, viz. systematically preserving the geographically stratified dialect lexicon, but they also provide a cultural and historical testimony of the everyday discourse practices of the dialect speakers in the early 20th century.

Both dictionaries elaborate in detail about their geographical scope in the introduction. As it is nearly impossible to distinguish dialect regions which are valid for every linguistic aspect of a particular dialect, the dictionaries rely on administrative borders. The WBD comprises the province of North Brabant in the Netherlands and the provinces of Antwerp and Flemish Brabant (including the Brussels region) in Belgium.[5] The data for the WLD come from the Netherlandic province of Limburg, the Belgian province of

1 Before Weijnen's call for collaboration, other scholars, including J. Schrijnen, J. van Ginneken and W. Roukens in the Netherlands and L. Grootaers and J.L. Pauwels in Belgium, already launched comparable initiatives on a smaller scale. This resulted in the NCDN having at its disposal a large collection of systematically collected dialectological (questionnaire) data from the Belgian and Netherlandic Limburgish dialects.

2 To avoid terminological confusion, the term 'Flemish' will be reserved for reference to the Flemish dialect area, in the west of the northern part of Belgium. In Dutch, 'Flemish' and 'Flanders' can also refer to the northern part of Belgium as a whole (i.e. the Dutch-speaking part, which also comprises the Brabantic and Limburgish dialect areas). We will, however, only refer to the latter region as 'the northern part of Belgium' or as 'the Dutch-speaking part of Belgium'.

3 Most of the databases of the general vocabulary part of the WBD and WLD are available online (http://dialect.ruhosting.nl/wbd/ and http://dialect.ruhosting.nl/wld/). In 2016, a new website for the WLD, which contains the databases for the entire dictionary was launched: http://e-wld.nl. A similar website will be completed for the WBD in December 2017.

4 See http://www.wvd.ugent.be/dsdd

5 Initially, the border between the Brabantic and Limburgish area was set, by the Brabantic editors, at the so-called 'ideal Gete line', a bundle of isoglosses that runs through Limburg and that was described by J.L. Pauwels & L. Morren in 1960. However, by the time the introduction to the WLD was published, the editors of both dictionaries had decided to use the administrative borders of the provinces instead (Weijnen, Goossens & Goossens 1983: 7-9).

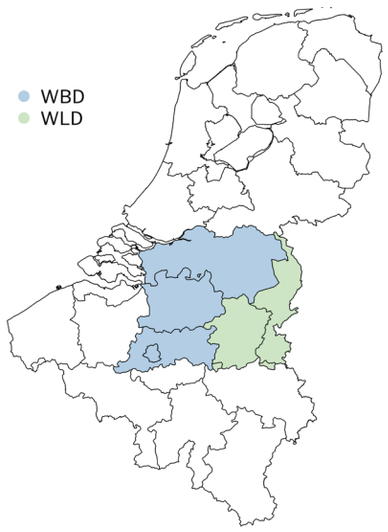


FIGURE 2.1
The Brabantic and Limburgish dialect area in the WBD and the WLD

Limburg and the north of the province of Liège.[6] Figure 2.1 shows the geography of the dialect regions relative to the rest of Belgium and the Netherlands.

The largest part of the lexical items in the dictionaries are elicited by means of the questionnaires distributed by the NCDN. Additionally, both dictionaries also use supplementary sources, like other questionnaires with a more limited geographical or thematic scope (like the data that were collected by Schrijnen, Van Ginneken & Verbeeten in 1914, questionnaires distributed by the Meertens Institute etc.), or local dialect dictionaries, like the dictionary of the dialect of Diepenbeek (published by J. Castermans in 2000). The relevant sources are listed at the beginning of each volume.

In our analyses, we only use data from the third part of the dictionaries, concerning the general vocabulary, for practical reasons. Additionally, general vocabulary is expected to be known by every dialect user, whereas knowledge of the specialized terminology collected in part 1 and 2, depends on the profession or social environment of the dialect speaker (Weijnen, Goossens & Goossens 1983: 6). Every part consists of a number of volumes. Each volume concerns a particular semantic field and is subdivided into several semantic subdomains. An overview of the 14 general vocabulary volumes of the WLD with some exemplary semantic subdomains, is provided in Table 2.1. The volumes are organized in four large parts, from human beings on their own, to people's relationship with and perspective on the outside world. The first four volumes concern mankind as a human being (III, 1). The second group of volumes focuses on domestic life (III, 2). The next three volumes deal with man as a social being (III, 3) and the final four volumes with the outside world (III, 4). The data included in each volume are highly comparable in both dictionaries.

6 Originally, a Germanic dialect, which bears similarity to the other Limburgish dialects, was spoken in the north of the province of Liège (Weijnen, Goossens & Goossens 1983: 7-9). A large part of this region also historically belonged to the Duchy of Limburg.

2.3 THE STRUCTURE OF THE DATABASES

As outlined above, we use the offline version of the databases of the WBD and WLD in this dissertation and we only focus on the general vocabulary. One separate tabular file is available per volume (i.e. semantic field) per dictionary. One row in these files represents one observation of a dialectal variant (word form) in a particular location for a specific question (or from another source). Of course, for each observation, the corresponding concept is available as well, making the files easy to use onomasiologically. The dataset also contains information about the source of the observation and the wording used in the questionnaires. The so-called 'kloekecode' for the location where the observation was recorded, is available as well. These codes form a unified way of referring to the locations in the Dutch language area. In the dictionaries, the 1962 version of the list of kloekecodes is used. This list was originally devised by L. Grootaers & G.G. Kloeke and later revised by W. Pée and P.J. Meertens.

Table 2.2 provides an example of the relevant parts of the dataset, taken from the semantic field of clothing & personal hygiene in the WBD. As the table indicates, the data can come from different sources, like the questionnaires of the NCDN (indicated as N + questionnaire number + year of distribution, e.g. N 86 (1981)), other large-scale questionnaires (e.g. ZND) and local dictionaries (e.g. Leuven Wb. 1, a dictionary for the dialect of Leuven). For every concept, a large set of records is available from different locations. Moreover, for some locations, more than one record is available and, sometimes, the same lexical item occurs twice in one place. This happens when more informants from the same location filled out the questionnaire (BRACELET in Aarschot), or when various questions are combined into one concept (UNDERVEST in Steenbergen). Furthermore, when lexical variants have a non-native origin, this is generally marked with a tag (*bracelet* for BRACELET is marked as French).

The dataset also contains information that will not be used in this dissertation. In setting up the volumes of the dictionaries, phonological variants were transcribed in a 'dutchified' form ('lexical item' in Table 2.2) using a relatively well-defined set of rules (like sound laws, see Weijnen, Goossens & Goossens 1983: 34-44). As the aim of this dissertation is to determine variation in the amount of lexical diversity per concept, we rely on these dutchified forms

volume		examples of semantic subdomains
III, 1.1	the human body	body parts, organs and their functions, the senses
III, 1.2	physical motions & health	posture, health and disease
III, 1.3	clothing & personal hygiene	clothing, hats, shoes, grooming, hygiene
III, 1.4	personality & feelings	personality traits, intellect, feelings, behaviour
III, 2.1	the house	rooms in the house, maintenance, silverware
III, 2.2	family & sexuality	kinship, getting married, having children, end of life
III, 2.3	food & drink	food, meals, cooking, fruits, vegetables, smoking
III, 3.1	society, school & education	transportation, education, the judicial system, language
III, 3.2	celebration & entertainment	local festivities, sports, children's games, the arts
III, 3.3	church & religion	the clergy, Christian religion, liturgy, the church building
III, 4.1	fauna: birds	birds of prey, birds in the woods, birds around the house
III, 4.2	fauna: other animals	insects, reptiles, fish, wild mammals
III, 4.3	flora	trees and shrubbery, wild plants and flowers, moss and fungi
III, 4.4	the physical & abstract world	quantities, weather phenomena, land and water, time

TABLE 2.1
Overview of volumes and exemplary semantic subdomains in the WLD

rather than taking into account phonological details. Other information that is available in the databases includes an identification number for most of the informants and additional remarks from the informants or lexicographers.

In the analyses, we only include observations that contain information about which concept the record belongs to, the dutchified form of the dialectal variant (i.e. the lexical item), the source of the observation and the location where the observation was recorded. How we use these observations to calculate variation in lexical diversity will be discussed in each chapter separately.

2.4 ADVANTAGES AND LIMITATIONS OF THE DATASET

Language research should be based primarily on **naturalistic** language data (in line with the usage-based approach of the Cognitive Linguistics paradigm; see for instance Bybee & Hopper 2001, Dabrowksa & Divjak 2015, Geeraerts 2010, Kemmer & Barlow 2000, Langacker 1988a, Tomasello 2001). In the databases of the WBD and WLD, this commitment is largely met due to the fact that most of the data are based on questionnaires filled in by common language users. The reliability of these data is safeguarded by the fact that the NCDN researchers re-elicited unclear lexical items that occurred only once for a particular concept (Weijnen & Van Bakel 1967: 31). They did this to ensure that the dialect user is truly relying on his/her knowledge of the local dialect, rather than providing a word that incidentally comes to mind, because (s)he does not know or remember a dialectal word form for the concept questioned.

The databases have three further advantages that are essential in order to meet the aims of this dissertation. First, measuring lexical diversity requires a dataset with an **onomasiological** perspective in order to compare *all* the lexical variants that are available to refer to a particular referent. Second, the dataset has to be **large** enough to investigate the influence of the explanatory variables on variation in the number of names for a concept in a reliable and valid way. Third, the influence of semantic features on variation in the amount of lexical diversity has been established in the

concept	question	lexical item (dutchified)	source	location	kloekecode
armband ‘bracelet’	band- of ringvormig, gewoonlijk metalen sieraad dat om de arm of pols gedragen wordt (armband, bracelet)	bracelet (fr.)	N 86 (1981)	Aarschot	P025p
armband ‘bracelet’	band- of ringvormig, gewoonlijk metalen sieraad dat om de arm of pols gedragen wordt (armband, bracelet)	bracelet (fr.)	N 86 (1981)	Aarschot	P025p
armband ‘bracelet’	band- of ringvormig, gewoonlijk metalen sieraad dat om de arm of pols gedragen wordt (armband, bracelet)	armband	N 86 (1981)	Aarschot	P025p
armband ‘bracelet’	band- of ringvormig, gewoonlijk metalen sieraad dat om de arm of pols gedragen wordt (armband, bracelet)	armband	N 86 (1981)	Halsteren	I078p
armband ‘bracelet’	band- of ringvormig, gewoonlijk metalen sieraad dat om de arm of pols gedragen wordt (armband, bracelet)	bracelet (fr.)	N 86 (1981)	Landen	P171p
armband ‘bracelet’	armband	armband	ZND 32 (1939)	Essen	K189p
armband ‘bracelet’	armband	armband	ZND 32 (1939)	Poppel	K196p
armband ‘bracelet’	armband	bracelet (fr.)	Leuven Wb. 1	Leuven	P088p
…	…	…	…	…	…
borstrok ‘undervest’	borstrok, warme onderkleding, gedragen over het hemd (borsrok, hemdrok, hemdsrok, hemsrok)	borstrok	N 02 (1960)	Attenhoven	P169p
borstrok ‘undervest’	borstrok, warme onderkleding, gedragen over het hemd (borsrok, hemdrok, hemdsrok, hemsrok)	slaaplijf	N 02 (1960)	Landen	P171p
borstrok ‘undervest’	borstrok, warme onderkleding, gedragen over het hemd (borsrok, hemdrok, hemdsrok, hemsrok)	borstrok	N 02 (1960)	Steenbergen	I057p
borstrok ‘undervest’	borstrok, onderkledingstuk dat over het hemd wordt gedragen [hemdrok, humperok, sjtoep, liefke, slaoplijf]	hemdrok	N 25 (1964)	Steenbergen	I057p
borstrok ‘undervest’	borstrok voor vrouwen	borstrok	N 25 (1964)	Steenbergen	I057p
borstrok ‘undervest’	borstrok voor mannen	hemdrok	N 25 (1964)	Steenbergen	I057p

TABLE 2.2
Excerpt of the relevant columns from the semantic field of clothing & personal hygiene in the WBD (volume III, 1.3)

semantic field of the human body in the Limburgish dialects (Geeraerts & Speelman 2010, Speelman & Geeraerts 2008), but the degree to which these features also affect lexical variation in **other semantic fields** and in **other datasets** has not yet been systematically investigated. To corroborate these findings, data is needed from other semantic fields than the human body and from more language varieties. The perspective chosen in this dissertation is to examine both the Limburgish and Brabantic dialect data to extend the scope to other semantic fields than the human body.

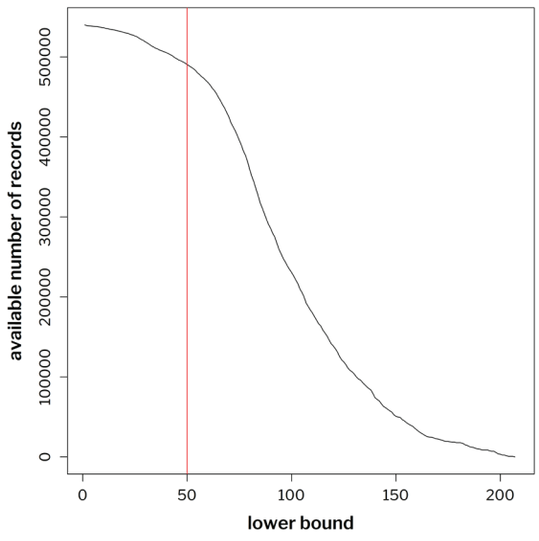


FIGURE 2.2
Available number of records (y-axis) in function of an increasingly higher lower bound (x-axis) in the general vocabulary part of the WLD

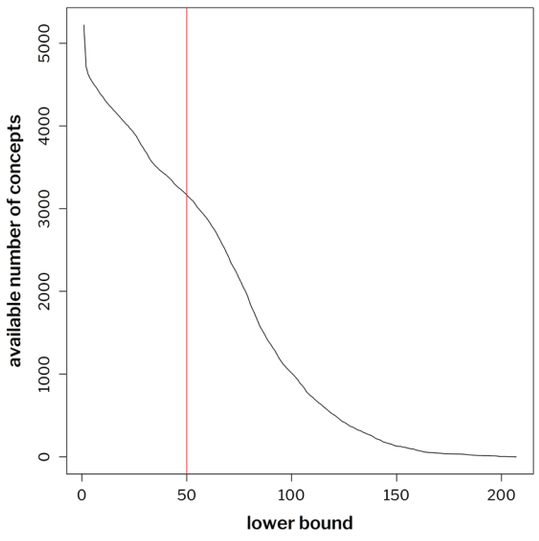


FIGURE 2.3
Available number of concepts (y-axis) in function of an increasingly higher lower bound (x-axis) in the general vocabulary part of the WLD

Both databases also have limitations which should be made explicit. First, when researching the variation a concept shows, it is important that data are available from the same set of locations for every concept investigated, because less variation will show up automatically when a smaller number of locations is taken into account. Since only the questionnaires of the NCDN were **systematically** distributed throughout the dialect areas and the other sources were used solely as supplementary material, we only use the data elicited with these questionnaires.

However, relying only on the questionnaire data does not ensure that dialect data from the same set of locations are available for every concept (De Vriend, Swanenberg & Van Hout 2007). Several reasons can be envisaged to explain this finding. Aside from a more limited distribution of the questionnaires than expected, data scarcity may also result from editorial interventions on the raw questionnaire data. For instance, if the lexicographers encountered a lexical item that can only be considered to be mistake by the informant, sometimes, the questions turned out to be not clear enough and new concepts were chosen to cover the answers (Weijnen & Van Bakel 1967: 30-33). This results in different quantities of data being available between the concepts in the database. A similar effect stems from the fact that the editors did not a priori decide which concepts would be included in the dictionary. Instead, concepts were selected on the basis of what the responses to the questionnaires revealed about the conceptual structure of the data. For instance, sometimes, the questionnaires were too fine-grained, which resulted in the elicitation of lexical items for one concept with more than one question. For example,

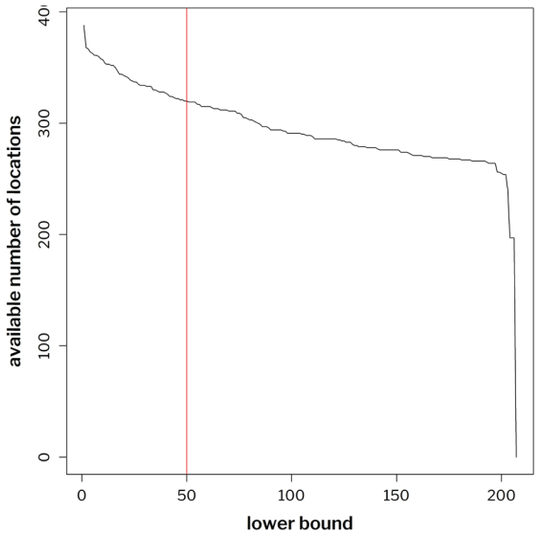


FIGURE 2.4
Available number of locations (y-axis) in function of an increasingly higher lower bound (x-axis) in the general vocabulary part of the WLD

instead of distinguishing between two questions like *shorts for boys* and *shorts (in general)*, the responses for these two questions are subsumed under the concept SHORTS. Finally, as will be discussed in chapter 3, some concepts may occur with responses from a smaller amount of locations, because they are less well-known. This can result in dialect users not returning a response to a question, either because they do not know the concept, or because they are unaware of or cannot remember the appropriate word in their local dialect.

In this dissertation, we provide a solution for such problems by setting a lower bound on the number of concepts for which data has to be available in each location and on the

number of locations for which data has to be present per concept (a similar approach is used in De Vriend, Swanenberg & Van Hout 2007).[7] More specifically, as a rule-of-thumb, we only include locations that have data for at least 50 concepts per semantic field and concepts that are occur in at least 50 locations per semantic field. Because these boundaries are imposed upon the data per volume and per dictionary, the locations that are included in the final dataset may differ between the semantic fields and dictionaries.

This lower bound may seem arbitrary, but it would be valid if a large amount of data (i.e. a large number of individual records, concepts and locations) remains available, while, at the same time, the lack of systematicity in the data is reduced as much as possible. Figures 2.2 - 2.4 show that 50 seems to be a legitimate cut-off point. The remaining number of records (on the y-axis) is plotted in function of an increasingly higher lower bound of concepts and locations in the general vocabulary part of the WLD (on the x-axis). Especially in Figures 2.2 and 2.3, it is clear that the slope of the black line shows a steep decrease just after the lower bound equal to 50 (indicated with the red line). The effect of the lower bound for the remaining number of locations is less strong (Figure 2.4). Only when a minimum of 200 concepts per location and 200 locations per concept has to be available, the slope decreases dramatically.

Another limitation has to do with the **origin of the data** in the dictionaries. The researcher has to be aware of the fact that (s)he is assuming a certain degree of homogeneity in the dataset at different levels, even though hardly any information is available about the background of certain aspects of the data. Firstly, the dictionaries aim to describe the dialect lexicon at the beginning of the 20th century, but in practice, not all the questionnaires were distributed at the same time.[8] As a result, the extent to which the data are homogeneous from a diachronic perspective is unclear. Crucially, it is not possible to uncover the degree of diachronic homogeneity, because each questionnaire elicits data on a different topic. Secondly, the dictionaries assume sociolinguistic homogeneity in the data, but the databases do not contain information about the social background of the informants. Furthermore, both dictionary projects have been led by changing editorial teams since the publication of the first volumes. This may also result in differences between the volumes, although the introductions do list a set of guidelines that have been followed. Finally, comparing the dictionaries directly is difficult, because differences may show up, due to the fact that they do not have identical editorial teams. For instance, in the WBD the variant *onjeklonje* for EAU DE COLOGNE is treated as a different lexical item than the lexeme *eau de cologne*. In the WLD, however, these variants are treated as phonological alternatives of the same dutchified word: they are both subsumed under the same lexical item *eau de cologne*.

These latter two problems (lack of homogeneity regarding the editorial team within and between the two dictionaries) are solved by making sure that in our case studies, all the operations are calculated for each semantic field and for each dictionary separately. We only directly interlink the dictionary data and execute calculations across dictionaries in chapter 6, in which we focus on plant name variation. This causes some problems that are to a certain extent related to differing editorial practices between the dictionaries (see the discussion in chapter 6).

However, we are not aware of a direct way of handling the other two disadvantages mentioned above (uncertainty about the degree of diachronic and sociolinguistic homogeneity). To cope with these problems, we conduct the analyses under the assumption that the data are relatively homogeneous and truly reflect the dialect lexicon of the early 20th century. At the same time, we remain alert for biases introduced by the dataset itself. We believe that the fact that we aggregate over a large amount of data accounts for a lot of the noise that may be present in the dataset. Evidence for this assumption comes from the fact that previous research using large-scale analyses of these dialect data yields sensible results (e.g. Cornips et al. 2016, Swanenberg 2000 and later, and the contributions in theme issue 20 of the journal *Taal en Tongval* (2007) concerning the usability of the southern Dutch dictionaries for linguistic research).

7 In chapter 6, this lower bound is not used for practical reasons.

8 The largest proportion of the general vocabulary data was elicited between 1960 and 1990 and a few supplementary questionnaires were distributed until 2003. These follow-up questionnaires often have a limited geographical scope. Questionnaire N104 (2000), for example, is a supplementary questionnaire that was predominantly distributed in the Netherlandic province of Limburg.

Case Studies – part 1

3. Revisiting lexical diversity in dialect data. The influence of semantic concept features beyond the human body

3.1 INTRODUCTION

Geeraerts & Speelman (2010) and Speelman & Geeraerts (2007, 2008) conducted pilot studies on the semantic field 'the human body' in the Dictionary of the Limburgish Dialects. The central aim of these studies was to determine whether lexical geographical heterogeneity (i.e. the amount of lexical and geographical variability a particular concept shows) is also influenced by features related to the prototype-theoretical organization of the lexicon. As explained in chapter 1, such a view of the lexicon takes for granted the fact that categories can show different degrees of onomasiological vagueness (i.e. non-discreteness) and onomasiological salience (i.e. non-equality). The pilot studies show that the degree to which a concept is vague or less salient, correlates with the number of lexical variants available for the concept in the Limburgish dialect data and the way these variants are heterogeneously scattered across the dialect area.

Additionally, the pilot studies take into account another, more traditional feature of lexical variation, viz. negative affect.[1] It is generally recognized that negatively connoted concepts, especially concepts that are prone to taboo, show a large amount of variation (Allan & Burridge 1988, 2006). Due to socio-cultural constraints (e.g. Grondelaers & Geeraerts 1998, Pizarro Pedraza 2015), language users often prefer to use euphemistic (or dysphemistic) expressions to refer to a tabooed concept. Sexual vocabulary, for instance, is characterized by lexical richness, due to the fact that lexical items for these highly negatively connoted concepts are prone to rapid language change, as these words quickly lose their euphemistic reading (Allan & Burridge 2006: 243). In the pilot studies, a more general variable than merely proneness to taboo was used, viz. negative affect. In practice, seven respondents were asked to rate the concepts of the semantic field of the human body for the degree to which "they expected people to attribute a negative connotation to the concept" (Speelman & Geeraerts 2008: 231). The results show that negative affect clearly serves as a source of lexical geographical heterogeneity across dialects as well.

However, some open questions remain. To what extent do the semantic features that were distinguished in the pilot studies also affect the variability of lexical diversity in other semantic fields than 'the human body'? Can we establish that they impact lexical variation in other dialect areas than the Limburgish region as well? Additionally, if other semantic fields are taken into account, we can expect to find that some fields are, in general, more prone to variation than others, due to socio-cultural conditions. Names for children's games, for instance, are notoriously diversified, due to the fact that children's language is highly imaginative and creative (Weijnen 1966: 336, Pickl 2013). In contrast, a field like religion is probably prone to less variability, because it is based on a highly standardized framework, accompanied by specific concepts and their corresponding names which are often directly borrowed into (dialectal) varieties. However, aside from socio-cultural features, factors relating to differences in cognitive processing and categorization may be at play as well. For instance, if we find less variation in the dialect lexicon for human body parts than in the words for a person's feelings, does this have to do with the fact that the former are generally more concrete (i.e. perceptible with the senses) and therefore more easily processed and entrenched?

The aim of this chapter is, therefore, to establish that the influence of the semantic concept features distinguished in the pilot studies is relevant for the dialect lexicon as a whole. In practice, we will apply a highly similar methodology to the one used in the pilot studies to investigate the

1 (Negative) affect is comparable to the concept of valence/evaluation/pleasantness, often measured at the level of the word rather than at the level of the concept, in psychometric research (Osgood, Suci & Tannenbaum 1957, also see e.g. Barrett & Russell 1999, Bradley & Lang 1999).

effect of the concept features in five other semantic fields, aside from the human body, of the WBD and WLD. By keeping the methodology as stable as possible, the results of this chapter are directly comparable to the results obtained in the pilot studies. Additionally, by adding other semantic fields to the analysis, we can inquire into features that may influence differences in the amount of lexical variation between semantic fields.

Before providing an overview of the results of this study in 3.5, we explain the theoretical framework and operationalization of the predictor variables in 3.2., the quantification of lexical geographical heterogeneity, the response variable in the analysis, in 3.3 and an overview of the statistical technique that is used in 3.4. Section 3.6 rounds of this chapter with a summary and discussion.

3.2 EXPLANATORY VARIABLES

In this chapter (and in chapter 4), we rely on the linguistic data in the databases to gauge the semantic characteristics, degree of onomasiological vagueness and salience, of the concepts under investigation under the assumption that semantic properties can be inferred by relying on the words that occur in the onomasiological range of the concepts under scrutiny. Additionally, self-reported data are used to quantify differences between the semantic fields included in the analysis and to determine the degree to which a concept is prone to affect. In chapter 6, these data will be complemented with referential data related to the environment of the language users.

3.2.1 Semantic field & dictionary

The main purpose of this chapter is to determine whether the results obtained in the pilot studies are also stable in other semantic fields and in two different dialect regions. Consequently, while the pilot studies focused on the semantic field of the human body in the Dictionary of the Limburgish Dialects, in this chapter, we also include data from a different dictionary, viz. the Dictionary of the Brabantic Dialects. We expect to find that the results are the same across both dictionaries. The dataset contains 3136 concepts in total. For the WLD, 1456 concepts are available, while 1680 concepts are included from the WBD. Additionally, the concept-based measurements are based on a larger number of observations in the WBD (328 320) than in the WLD (204 307). These differences may be related to the fact that the surface of the Brabantic dialect area is larger, contains a larger number of locations (408 versus 252 in the WLD) and has a more dense distribution (i.e. the locations in the Limburgish area are, generally, geographically more distant from each other). As explained in chapter 2, we use a minimum number of 50 places per concepts for the concept to be included in the dataset. It is possible that this lower bound of 50 is more quickly attained in the Brabantic data due to the fact that the geographical region is larger. The concept-based measurements are based on 532 627 observations in total, which amounts to about 170 tokens per concept on average.

We also take into account other semantic fields organized along two dimensions. First, we include a social dimension, viz. whether a semantic field is universal or related to local or supralocal societal phenomena. The second, semantic, dimension, concerns the average amount of concrete or abstract concepts in a semantic field. This allows us to examine to what extent differences that may show up between semantic fields can be explained by these features.

The first dimension is included following the results obtained by Pickl (2013). He showed that in dialect data from Bavarian Swabia, in the south of Germany, the chance of finding lexical levelling (i.e. a low degree of geographical heterogeneity) is higher for concepts relating to semantic fields that are socially relevant on a large scale, rather than locally bound. To select semantic fields that differ along this first dimension, we rely on the organization of the dictionaries into sections and volumes. As explained in 2.2, the general vocabulary part of the dictionaries consist of four sections: (1) man as an individual, (2) domestic life, (3) community life and (4) the world versus man, which are organized along an expansional axis from people on their own, to people's relationship with the outside world. Each section is further subdivided into three or four volumes. We selected one volume from sections 1, 2 and 3, but excluded data from section 4, which contains the lexicon of the world outside of man: plants, birds and other wild animals, and the physical and abstract world (e.g. weather phenomena, time and space). However, chapter 6 will inquire in detail into lexical variation in the Flora volume of section 4.

The second dimension lines up with research in psycholinguistics, which shows that concrete words are easier to process and to recognise than more abstract words (Gorman 1961, Hargis & Gickling 1978, Paivio 1986). As a result, the advantages in the processing of words for concrete concept may correlate with a larger degree of psychological entrenchment and, thus, a smaller amount of lexical variation in the dialect data under scrutiny. To determine the degree to which a semantic field contains more concrete or abstract concepts, we rely on word ratings for 30 000 Dutch words collected by Brysbaert et al. (2014). The degree of concreteness of a word is defined as "the degree to which a concept denoted by a word refers to a perceptible entity" (ibid.: 81), i.e. is based on experience, while an abstract word refers to a concept that is not perceptible with the senses

and language-based. We use the word ratings obtained by Brysbaert and colleagues to calculate the mean concreteness rating for each semantic field in the WLD.[2] These ratings range from 1 (very abstract/language-based) to 5 (very concrete/experience-based). Then, we select three semantic fields, one for each section that we use from the dictionaries, with a low value for mean concreteness, i.e. semantic fields that mostly contain abstract concepts, and three semantic fields, again one for each section, with a high value for mean concreteness, i.e. semantic fields that mostly contain concrete concepts. Importantly, the semantic fields that we use for the individual-based category (from section 1) mostly contain universal concepts. Although they are not related to societal phenomena, but to man as an individual, they are universal and salient for every living person (Lakoff & Johnson 1980). Table 3.1 shows an overview of all the semantic fields per section and the mean concreteness rating per semantic field.[3] Table 3.2 shows the semantic fields that are included in the analysis, organized along the two dimensions.[4] It also contains the number of concepts available per semantic field. The semantic field of the human body contains body parts and descriptions of body types, like HEAD, KNEE, FOOT and CORPULENT. This field also contains 20 concepts that are explicitly listed as jocular. The semantic field 'the house' includes concepts relating to objects and activities in and around the house, such as CUTLERY, TYPES OF POTS, CLEANING UP, WASHING and DOING DISHES. The semantic field of celebration & entertainment consists of sports & (children's) games (e.g. concepts related to football, card games and to play marbles), celebrations (e.g. CARNIVAL and other calendar-bound events) and concepts relating to the arts (e.g. SCULPTOR). In the semantic field of personality & feelings, a few larger groups are distinguished: concepts relating to (temporary) feelings (e.g. ANGER), non-temporary personality traits (e.g. TO BE SHY), types of behaviour (e.g. HASTY) and memory & thinking (e.g. TO INFORM, TO LEARN SOMETHING). The field of family & sexuality contains concepts relating to stages of life and death (and corresponding rituals), like baptism, marriage, and death and burial. The field of society, school & education, finally, concerns concepts relating to societal organisation (e.g. concepts concerning the police, war and defence), man's relation to society (e.g. concepts for language and communication), schooling and different types of transportation (eg. by car or other types of vehicles (e.g. LANDAU, JALOPY), by train, by air). Appendix 3.1 contains an overview of all the subsections per semantic field and example concepts.

3.2.2 Onomasiological vagueness

Background

Research on prototype effects in the lexicon has shown that the semasiological structure of individual words is characterized by non-equality, i.e. differences regarding the degree of membership, and non-discreteness, fuzziness at the edges of a category (see chapter 1). Prototype theory and its extensions are innovative because they show that the structure of categories, like birds or fruits, is characterized by variation in structural weight and by fuzziness from a *semasiological* perspective (Geeraerts 2010: 199-200). However, similar effects can also be found on the onomasiological level, i.e. within a semantic field (Geeraerts, Grondelaers & Bakema 1994).

Onomasiological vagueness concerns the degree of fuzziness at the borders of concepts belonging to a particular semantic field.[5] For example, for a non-specialist, the demarcation between a cactus (Cactaceae) and a spurge that grows in a drought habitat (e.g. Euphorbia lactea) is vague. In practice, both plants are often referred to as cacti, because perceptually, they have the prototypical shape and thorns typical for cacti and because they can survive in a drought environment (see Figure 3.1). However, biologically, they belong to different families. As Rosch (1978: 29) argues, the attributes that are perceived by a language user, are dependent on their functional needs, like the available category system and their physical and social environment. A layman has a cultural-linguistic category CACTUS, which is prototypically represented by a green, thorny, succulent plant that typically grows in a drought habitat. The category SPURGE, however, is far less familiar. Crucially, the *degree* of

2 The data from Bysbaert et al. (2014) are collected at the level of the word in standard Dutch. Following the authors, we assume that these standard Dutch variants reflect the degree of concreteness of the concept designated by the concept names to a great extent. We solely rely on concept names that are identical to the words available in the concreteness data. Although this may influence the mean values to some extent (e.g. through polysemy), we assume that we account for this possible source of noise in the data by aggregating over all the concepts in a particular semantic field.

3 Even though the semantic field of church and religion shows the smallest value for mean concreteness in section 3, we did not use it in this chapter because, as chapter 5 will show, this semantic field is characterized by very little geographical variability and a high amount of lexical standardisation. Additionally, we also use affect as a predictor variable in this study, but analyses of the mean valence per semantic field, on the basis of ratings collected by Moors et al. (2013), indicate that very few concepts from this field are prone to affect. Furthermore, concepts belonging to this field are less familiar for the raters on who we rely for our measure of affect-sensitivity (i.e. people living today), due to the fact that the influence of the church has decreased over the last decades (see chapter 5).

4 The data for the semantic field of family & sexuality from the WBD are unavailable.

5 Additionally, it can also be interpreted as fuzziness at the edges between two semantic fields (see Geeraerts et al. 1994 and chapter 1).

section	semantic field	proportion of concepts found in Brysbaert et al. (2014)	mean concreteness
1. man as an individual	the human body	0.590	4.390
	physical motions & health	0.612	3.677
	clothing & personal hygiene	0.208	4.316
	personality & feelings	0.579	2.347
2. domestic life	the house	0.449	4.345
	family & sexuality	0.488	3.359
	food & drink	0.487	3.967
3. community life	society, school & education	0.580	3.260
	celebration & entertainment	0.193	3.772
	church & religion	0.204	2.988
4. the world versus man	fauna: birds	0.347	4.051
	fauna: other animals	0.382	4.453
	flora	0.222	4.207
	the physical & abstract world	0.237	3.755

TABLE 3.1
Proportion of available concepts and mean concreteness on the basis of Brysbaert et al. (2014) in the four sections of part 3 (general vocabulary) in the WLD

	concrete		abstract	
	semantic field	**N**	**semantic field**	**N**
individual	the human body	361	personality & feelings	703
locally-bound	the house	508	family & sexuality (WLD only)	119
societal	celebration & entertainment	471	society, school & education	974

TABLE 3.2
Semantic fields used in the study organized along the two dimensions of interest and number of concepts per semantic field

FIGURE 3.1
On the left, a prototypical member of the cactus family (Opuntia ficus-indica); on the right, a Euphorbia lactea.[6]

vagueness of a category can differ between concepts: for a botanist, the distinction between cacti and these types of spurges is much more salient.[7]

As the cactus example shows, to empirically determine whether specific concepts are vague towards each other, the semasiological range of application of the words used for the concepts can be considered (see Geeraerts et al. 1994: 122-140, which also provides an overview of an alternative operationalization of onomasiological non-discreteness, viz. by means of definitional criteria, and of structuralist approaches to onomasiological vagueness). More specifically, if a particular lexical item, like *cactus*, is used for both a plant belonging to the Cactaceae family and for a species of Euphorbia, this serves as an indication that the two concepts cannot be clearly demarcated from each other in the speech community. Importantly, however, lexical items or constructions that can be used for more than one concept, can also be polysemous, rather than vague. In this case, the concepts to which the items refer are clearly distinct and not vague towards each other from an onomasiological perspective. For instance, the word *port* can be used for a type of drink and for a harbour, concepts that are onomasiologically unrelated. However, determining stable and congruent tests to distinguish vagueness from polysemy is problematic (Geeraerts 1993, 2015). Research in cognitive semantics, which takes for granted a prototype-theoretical view of language and rejects the distinction between linguistic and real-world knowledge, has instead argued that polysemy is a form of categorization and that the distinction between vagueness and polysemy is continuous, dynamic and determined by contextual and real-world knowledge (for an overview, see Lewandowska-Tomaszczyk 2007 and Gries 2015).

Operationalization

Following the pilot studies, we take into account the semasiological range of the words used for a particular concept to determine the degree to which the concept is onomasiologically vague. More specifically, we rely on the amount of lexical items used for the concept that occur for other concepts as well. For instance, the WBD contains two concepts, VROUW DIE GRAAG KWAAD SPREEKT 'woman who likes to gossip' and KWAADSPREKER 'person who gossips', that can be expected to be vague towards each other, because they mainly differ along the dimension of gender (female or unspecified). The onomasiological vagueness of the concepts towards each other is reflected by the fact that there

6 The pictures were downloaded from *The Europeana collections*, an online repository for cultural heritage (http://www.europeana.eu, Accessed on 19 August 2017).

7 This is, for instance, reflected by the existence of webpages like *Cacti or not?*, which is dedicated to distinguishing cacti from other succulents (https://cactiguide.com/cactiornot/, Accessed on 19 August 2017).

are 15 lexical items that occur for both concepts, including *klapekster*, *klapei*, *kwaadspreker* and *vuil tong*.[8] However, the concepts are, at the same time, clearly distinguishable from each other, because some dialect speakers explicitly incorporate the gender dimension for reference to the female concept (e.g. *kletswijf*, *kwaadspreekster*, *roddeltante*). A higher degree of onomasiological vagueness is expected to correlate positively with the amount of lexical geographical heterogeneity a concept shows, because, for vague concepts, the chance that dialect speakers from different locations all make the same demarcation choice is smaller. As indicated by the pilot studies, this results in a larger amount of lexical variation for the vaguer concepts (also see Pickl 2013).

The operationalization of onomasiological vagueness that we use, **lexical non-uniqueness**, was also used in the pilot studies on which this chapters builds. This measure calculates, for each concept, how often a lexical item belonging to the concept, also occurs in the database to refer to other concepts. Importantly, rather than calculating the degree of non-uniqueness between pairs of concepts, the measure is calculated per concept, to gauge the onomasiological vagueness of the target concept as a whole. If, for instance, a single lexical item of the target concept occurs for two alternative concepts, the measure takes a value of 2. To reduce the chance that a lexical item used for more than one concept is actually polysemous, we calculate non-uniqueness per semantic field. The variable ranges from 0 to 257, with mean 16.76 and standard deviation 25.15. It differs significantly between the dictionaries ($t = -10.522$, $df = 2737.1$, $p < 0.001$).[9] The mean for lexical non-uniqueness in the WLD is 11.95 (i.e. on average, per concept, about 12 lexical items are also used as a word for other concepts), with standard deviation 17.08. In the WBD, the mean is 20.94, with standard deviation 29.84. As it is unlikely that people from the Brabantic area generally consider concepts to be more vague than people from the Limburgish area, this finding is probably related to differences in the creation process of the two dictionaries.[10] Concepts with a high degree of lexical non uniqueness are, in the WBD, EZELACHTIG PERSOON 'a Simple Simon' (257), and, in the WLD, DEUGNIET 'rascal' (127). Concepts with a value of zero include VLIEGTUIG 'airplane' and TELEFOON 'telephone' in the WBD and BEELDHOUWER 'sculptor' and ZENUW 'nerve' in the WLD.

3.2.3 Lack of onomasiological salience

Background

The concept of onomasiological salience was introduced in Geeraerts et al. (1994), who relate it to the basic-level hypothesis (Berlin 1972, 1978, Berlin, Breedlove & Raven 1973). This hypothesis is based on the fact that, cross-linguistically, folk biological classifications consist of a limited set of taxonomical levels, which reflect the degree of onomasiological salience of the organisms involved. First, referents with a high degree of onomasiological salience, viz. referents that can be considered generic taxa (e.g. OAK, ROBIN), constitute the core of any folk biological organisation and, thus, the basic level: "[a]t this rank, both plants and animals appear perceptually most distinct to the human classifier, and these differences in morphology and behaviour virtually 'cry out to be named'" (Berlin 1978: 24). Second, the onomasiological salience of the basic level is reflected by the high frequency of the name situated at the basic level. Rosch et al. (1976), for instance, show that, in a free naming task, participants almost exclusively rely on the name associated with the basic category to refer to objects across nine taxonomies (viz. tree, bird, fish, fruit, musical instrument, tool, clothing, furniture and vehicle). Third, the basic level tends to be named with primary lexemes, i.e. "unique 'single-words' that can be said to be semantically unitary and linguistically distinct" (Berlin 1972: 54). For subordinate categories, binomial secondary lexemes, like *jack oak* (a kind of OAK), are generally used. Consequently, the basic-level hypothesis also contains predictions that connect the degree of onomasiological salience of a referent to *formal* properties of the names with which it occurs: the high onomasiological salience of the basic-level categories is further reflected by the fact that simple, single-word forms are used to refer to these concepts.

Geeraerts et al. (1994) argue that the basic-level hypothesis is problematic for two reasons when applied to other types of categories, like artefacts (in their case: clothing). Firstly, the hypothesis presupposes a neat taxonomical organization of the lexicon, because it is based on inclusion relationships: "each category is entirely included within one other category (unless it is the highest level category) but it is not exhaustive of that more inclusive category" (Rosch 1978: 30). However, clothing items like BROEKROK 'culottes' and DAMESKLEDINGSTUK 'woman's garment, item of clothing typically or exclusively worn by women' pose problems in such a view, as they are both difficult to place in a taxonomy

8 The total number of unique types is 45 for VROUW DIE GRAAG KWAAD SPREEKT and 63 for KWAADSPREKER.

9 In the remainder of this section, we use two-tailed unpaired t-tests to determine whether significant differences in the distribution of the independent variables between the dictionaries can be found (unless mentioned otherwise).

10 Because the semantic field of family & sexuality from the WBD is unavailable, the difference between the dictionaries may be related to the fact that concepts belonging to this field are more vague. However, if the distribution of lexical non-uniqueness in the two dictionaries is compared for only the five fields that occur in both dictionaries, the difference only becomes slightly smaller (WLD: mean = 12.50, sd = 17.54) and remains significant ($t = -9.6751$, $df = 2791.9$, $p < 0.001$).

in which SKIRTS and PANTS form the basic level. The authors argue that the lexicon seems to be organized more in the form of overlapping taxonomies that are all based on different dimensions. While a SKIRT and PANTS may be considered as functional gestalts, the concept DAMESKLEDINGSTUK reflects a different organizational principle of the lexicon, viz. along the dimension of gender. Secondly and more importantly, Geeraerts and colleagues show that, for artefacts like clothing items, onomasiological typicality exists between categories *on the same level of a taxonomical hierarchy* as well. For this reason, they propose to take into account a generalized notion of onomasiological salience, which they relate to Langacker's notion of entrenchment. Crucially, this approach allows them to show that differences in onomasiological salience, both between and within taxonomical levels, correlate with naming preferences, including the fact that concepts that are more entrenched from an onomasiological perspective, are more likely to be named with simplex forms (ibid.: 178-187).

Operationalization

To gauge the degree to which a concept is salient, we rely on several measures. The first two were also used in the pilot studies. The latter two serve as alternative ways to measure the degree of onomasiological salience of a concept. Following the pilot studies, we expect that "if a concept is less common, it is communicatively less prominent, and the possibility (or perhaps also the necessity) for standardization is more restricted" (Geeraerts & Speelman 2010: 27). Consequently, the hypothesis we aim to prove with these measures is that less salient concepts show a higher amount of variation (also see Szelid & Geeraerts 2008).

First, following the pilot studies, we rely on the **relative number of multi-word expressions** (MWE's) that occur per concept as a measure of lack of salience. The rationale for using this operationalization is two-fold. First, it relates to the predictions of basic-level theory regarding the formal properties of names for onomasiologically salient concepts. As explained above, these concepts tend to be named with simplex forms, while less salient concepts are often referred to with more complex lexical items. A second reason is that the dataset also contains hesitant, periphrastic expressions that seem to have been elicited because the respondents were not familiar with either the concept itself or with the dialect name for the concept. Such periphrastic responses can, for example, be found in the database of the WBD for the concept ONRUSTIG PERSOON 'restless person, a fidget'. About half of the observations for this concept, are a one-word lexical item, like *ongedurige*, *woelewater* or *zenuwpil*. The other half of the tokens, however, consist of periphrastic constructions, like *levendige kwiek* 'lively chap', *onrustige mieter* 'restless character', or *je kan een ei in zijn kont gaar koken* 'you can cook an egg in his bottom'.

We operationalize the proportion of multi-word responses per concept by dividing the total number of multi-word tokens[11] that occur per concept by the total number of observations. The variable ranges from 0 to 1, with mean = 0.124 and standard deviation = 0.223. It does not significantly differ between the dictionaries ($t = -1.0072$, $df = 3064.1$, $p > 0.1$). We expect a positive correlation between this variable and the dependent variable, lexical geographical heterogeneity (see below). Concepts with a value equal to 1 for this variable are TER BEGRAFENIS GAAN 'to go to a funeral' in the WLD and MEINEED PLEGEN 'to commit perjury' in the WBD. Concepts with a value of 0 include AUTO 'car' in the WLD and KNIE 'knee' in the WBD.

A second operationalization of lack of salience that was also used in the pilot studies takes into account the **proportion of missing places**, i.e. the proportion of locations for which no responses were provided for the concept.[12] The assumption is that for lesser known concepts, more respondents either did not know the concept or the dialectal name for the concept, which results in a larger proportion of missing locations. However, the interpretation of this variable is ambiguous, because a higher number of missing locations for the concept can also cause a smaller amount of lexical variation, as less data is available for the concept. The latter effect was found for this variable in the pilot studies: in these studies, number of missing places showed a positive correlation with the amount of variation per concept and the validity of the measure was already problematized.

This variable is calculated as follows. First, we obtain, for each semantic field in each dictionary, the total number of locations that are available. Then, we subtract the amount of locations with data per concept from this number. Finally, the resulting figure is again divided by the total number of locations per semantic field per dictionary. The variable ranges from 0 to 0.863 and differs significantly between the WBD and WLD ($t = -13.856$, $df = 3050.8$, $p < 0.001$). In the WLD, the mean is equal to 0.404, with standard deviation = 0.201. The mean value in the Brabantic data is equal to 0.503, with standard deviation 0.197. Concepts with a high

11 In this dissertation, the words 'token', 'observation' and 'response' are used interchangeably. They all refer to a single response of a particular respondent in a particular location to a specific question.

12 Alternatively, the number of responses per concept could also serve as a measure of salience, but since the number of respondents differs per location and as a higher number of responses is probably also available for affect-sensitive concepts (e.g. through euphemism), this measure seems less reliable.

proportion of missing places are ROOS VAN DE SCHIETSCHIJF 'bull's eye' (0.803) and WINKEL DRIJVEN 'to run a shop' in the WBD (0.863).

The first measure of lack of salience that was not included in the pilot studies is the **proportion of hapax legomena** per concept. Theoretically, using this feature for lack of salience depends on the rationale behind the operationalization of onomasiological salience in Geeraerts et al. (1994). More specifically, a concept, say COFFEE, is considered onomasiologically salient if a large proportion of the lexical items used for the concept are a unique name for the concept under scrutiny (e.g. the word *coffee*). A concept, like CAFFÈ LATTE, is less onomasiologically salient because, when all the names for the latter concept are taken into account, a higher proportion of these names can be considered as less typical for the concept. These non-typical names can include hyperonyms (e.g. *coffee*), hyponyms (e.g. *pumpkin spice latte*) or cohyponyms (e.g. *cappuccino*). Consequently, we can assume that a larger amount of non-typical names per concept indicates that the concept is not salient (we only take into account the non-typical names that occur once). Crucially, by using the proportion of hapaxes per concept we also take into account the fact that if a particular lexical item occurs only once for a particular concept in an entire dialect area, the chance that the dialect speaker made up this response on-the-spot is higher.[13] As explained above, these types of hesitant expressions sometimes occur in the database, when the dialect speaker does not know or cannot recall the name for the concept that is elicited, or when (s)he is unfamiliar with the concept itself. Consequently, the higher the proportion of hapax legomena for a particular concept, the higher the chance that the concept is less salient.[14]

The proportion of hapaxes per concept is calculated by dividing the total number of hapax legomena by the total number of tokens per concept. It ranges from 0 to 0.794, with mean = 0.088 and standard deviation = 0.089. It does not significantly differ between the dictionaries (t = -1.481, df = 3097.7, p > 0.1). Concepts with a high proportion of hapaxes are VERSCHILLENDE KNIKKERSPELEN 'various games of marbles' in the WLD (0.794) and GELUIDLOOS EEN WIND LATEN 'to let off a fart silently' in the WBD (0.635). Concepts without any hapaxes include BLOED 'blood' in both dictionaries, FAKKEL 'torch' in the WLD and EETLEPEL 'spoon' in the WBD.

The last operationalization of the degree of salience of a concept is based on the **prevalence** value of the name that is used in the dictionary to describe the concept. The prevalence data that are used were collected by Keuleers et al. (2015) in a large-scale online lexical decision experiment completed by over 365 000 participants from Belgium and the Netherlands. In this study, word prevalence is defined as "the proportion of a population knowing a particular word" (ibid.: 5). We automatically link the prevalence data, collected at the word-level, to the names for the concepts as they are available in the two dictionaries, under the assumption that the higher the prevalence value of a concept name, the more onomasiologically salient the concept is. The concept names that occur in the prevalence data all have a high prevalence value (the minimum z-score ranges from -1.243 to 1.960 with mean = 1.603 and sd = 0.466) However, for only 42.2% percent of the concepts from the dictionaries (1813 out of 3136 concepts), a prevalence value is available. On the one hand, this may have to do with the fact that the words that were used in the prevalence study are obtained from dictionaries and large-scale corpora, which indicates that, although many words with very low frequencies are included in the prevalence study (ibid.: 4), they are probably frequent enough to occur in linguistic corpus data. This is not always the case for the data in the dictionaries, as some of the concepts have become almost obsolete (e.g. WAMBUIS 'jerkin') or are limited to very colloquial speech (like specific types of children's games, e.g. BIKKELEN 'to play knucklebones'). On the other hand, however, many of the concepts that are not found in the prevalence data (viz. 61.9%), are listed in the dictionaries in the form of a multi-word expression (e.g. LASTIG (WERK); MUIS VAN DE HAND; PRATEN, KLETSEN). Since multi-word expressions are not included in the prevalence study, these concepts are not automatically found. However, according to basic-level theory, the fact that they are listed in a multi-word form already indicates that many of them, like MUIS VAN DE HAND, are less salient as well.

Consequently, to obtain a prevalence value for *all* the concepts in the dictionaries, instead of just for the ones that are automatically found in the data from Keuleers and colleagues, we also include a **binary operationalization** of concept prevalence in the analyses. In practice, this variable measures whether we are *certain* that a concept is preva-

13 As explained in chapter 2, the lexicographers of the dictionaries made sure to re-elicit data if unclear responses only occurred once in a geographical region.

14 To some extent, using the proportion of hapaxes as an explanatory variable is circular, because, as will be explained below, part of the response variable relies on the number of unique types that is available per concept, which is per definition larger for concepts with a larger number of hapaxes, given that the number of tokens is the same. However, relying on the proportion of hapaxes per concept, serves as a relatively direct way to turn around the rationale from Geeraerts et al. (1994) without the necessity of relying on formal properties of the lexical items used for the concept: we quantify the extent to which the lexical items are probably not unique for the concept, which is apparent from the fact that they occur only once. Furthermore, the other aspect of the composite response variable, geographical fragmentation, takes a profile-based approach to quantifying lexical diversity. As a result, the direct correlation between proportion of hapaxes and the composite response variable is reduced.

numeric	N available	minimum	mean	maximum	sd
min. of z-score BE & NL	1813	-1.243	1.603	1.960	0.466
mean of z-score BE & NL	1813	-0.889	1.677	1.960	0.403
max. of z-score BE & NL	1813	-0.566	1.751	1.960	0.357
categorical	**N available**	**prevalent**		**missing / not prevalent**	
prevalence binary	3136 (all data)	N = 1714 concepts (mean z-score BE & NL > 0.85)		N = 1422 concepts (of which 1323 missing)	

TABLE 3.3
Overview of different operationalizations of prevalence, viz. using the minimum, mean and maximum of the z-scores for prevalence in Belgium and the Netherlands, and using the binary operationalization of prevalence

lent.[15] The concepts for which the mean of the standardized z-scores for Belgium and the Netherlands is higher than 0.85 (this corresponds to 80.2% of the population in Belgium and the Netherlands, on average, knowing the concept name) are categorized as 'prevalent' (N = 1714). Concepts with the mean of the standardized prevalence scores equal to or smaller than 0.85 (N = 99) and concepts that are not available in the prevalence data, are categorized as 'missing/not prevalent'. The distribution of the prevalence variables is presented in Table 3.3. As the numerical standardized prevalence values do not reach significance in a multifactorial environment[16], possibly due to the fact that not enough data are available for these variables, we will only discuss the binary operationalization of the prevalence variable in the remainder of this chapter. The amount of prevalent (versus missing/not prevalent) concepts differs significantly between the dictionaries ($X^2 = 7.9464$, df = 1, $p < 0.01$; Cramer's V = 0.051). The proportion of prevalent concepts in the WLD is 0.598 (871 out of 1456 concepts), while it is a little lower in the WBD (0.547: 920 out of 1680 concepts).[17] Prevalent concepts include, in the WLD, OPSCHEPPEN 'to brag' and SPROOKJE 'fairy tale' and, in the WBD, HEMEL 'heaven' and SNURKEN 'to snore'. Missing / not prevalent concepts are UNSTER 'weighbeam' in the WLD and TIEND 'tithe' in the WBD.

3.2.4 Affect

While the proneness to affect of a concept is not directly related to a prototype-theoretical view of the lexicon, it is congruent with the maximalist perspective of meaning of the Cognitive Linguistics movement. As psychological research indicates that language users have clear positive or negative associations with words denoting particular concepts (Osgood, Suci & Tannenbaum 1957), affect constitutes one concept-related aspect of such an encyclopaedic view on category structure. In contrast with the measures for vagueness and lack of salience, affect is measured using information external to the dictionaries. First, a database of psychometric ratings for the valence of words is consulted (Moors et al. 2013). Additionally, as only 22.03% percent of the word lemmas in the WBD and WLD occur in this database, the analysis will predominantly focus on additional affect ratings collected with a forced-choice task.

Mean valence

Moors et al. (2013) collected ratings for affect (valence, arousal and dominance) and age of acquisition for 4300 Dutch words.[18] Each participant rated the entire list of words for one of these dimensions. However, we only use the valence ratings, as we are mostly interested in the degree to which a concept shows positive or negative connotations. Participants in the valence condition were asked to rate whether the words refer to something that is positive/

15 The prevalence data are available separately for Belgium and the Netherlands. We compared the influence of the minimum, maximum and mean of the values obtained in the two countries. All three variables are highly correlated (0.822 < Spearman's rho < 0.975, p < 0.001), but because using the mean z-scores has the largest impact on the response variable, we rely on the mean z-score for prevalence in Belgium and the Netherlands to calculate the binary predictor variable.

16 Comparing the impact of the numerical prevalence scores on the response variable with bivariate correlation tests, indicates that these variables have the expected effect (-0.199 < Spearman's rho < -0.192; all p < 0.001): the higher the prevalence value of a concept name, i.e. the better known the concept is, the less variation it shows.

17 If the concepts belonging to the field of family & sexuality, which are unavailable in the WBD, are excluded, the proportion of prevalent concepts in the WLD (0.607; $X^2 = 10.344$, df = 1, p < 0.01) and Cramer's V (0.059) are slightly higher. This indicates that the number of truly prevalent concepts in the field of family & sexuality is somewhat low in comparison to the other semantic fields.

18 The database is available online at http://crr.ugent.be/programs-data/word-ratings (Accessed on 20 August 2017).

pleasant rather than negative/unpleasant, by using a scale that ranges from 1 (very negative/unpleasant) to 7 (very positive/pleasant).[19] We automatically linked the words for the concepts in the WBD and WLD to the mean value of these valence ratings provided by Moors and colleagues. However, only 691 out of the 3136 concept names in the database (22.03%) occur literally in the psychometric valence data. This is probably related to the fact that the researchers explicitly excluded words that are not frequent in written corpora, that are obsolete or that are unknown in either Belgium or the Netherlands (ibid.: 171). The lexical dialect data, however, does contain a large number of concepts of this type. Additionally, similar to the prevalence data collected by Keuleers et al. (2015), the items that were rated never consisted of more than one word, in contrast with the names for many of the concepts in the Brabantic and Limburgish data. The distribution of mean valence in the dialect data is as follows: minimum = 1.5, mean = 3.945, maximum = 6.53, standard deviation = 1.081. The distribution of this variable does not differ significantly between the dictionaries ($t = 0.2275$, $df = 628.97$, $p > 0.1$). Concepts with the lowest value for valence (i.e. very negative concepts) are MISKRAAM 'miscarriage' and OORLOG 'war' (mean valence = 1.5). Concepts with a high value are LIEFDE 'love' (6.53), VREUGDE 'joy' (6.5) and VRIEND 'friend' (6.5). Examples of neutral concepts (i.e. mean valence = 4) are SLUIS 'lock' and VOOGD 'guardian'.

Forced-choice task

As for only 22.03% percent of the concepts in the database, valence ratings are available, we rely on a forced-choice task to collect affect ratings for the entire dataset. In contrast with the pilot studies, which only took into account the influence of *negative* affect on lexical variation, we also verified the degree to which positive concepts affect the amount of variation that a concept shows. The lexical richness of taboo-like concepts is generally accepted, but we also expect to find more variation for positive concepts like kinship terms or children's games (Pickl 2013).

The procedure for coding for affect was as follows. The raters received a tabular file containing the name of the concepts that are used in this case-study, organized per semantic field. Before distributing the file, we linked the concepts in the WBD to the concepts in the WLD to ensure that identical concepts receive the same affect rating.[20] In total, the file contains 1935 concepts.[21] The file also provides information about the subsection in the dictionaries to which the question belongs (e.g. ONNOZELE-KINDERENDAG 'Holy Innocents' Day' is listed under calendar-bound practices, which is subsumed under festivities and practices in the semantic field of celebration & entertainment) and a definition of the concept.[22] The raters also received a document with instructions. They were asked to use their best judgement in deciding whether a certain concept has a connotation. They could choose between four values: negative connotation, positive connotation, neutral concept or uncertain. Five raters coded the concepts in the database for affect, but only three of them completed the ratings for every semantic field. Four raters were early-stage researchers at the Department of Linguistics of the KU Leuven. One participant was an older, highly educated female, external to the department.

Although we could have used numerical scales to collect these ratings, we deliberately used a three-way forced-choice task for two reasons. First, preliminary analyses indicated that the mean valence ratings obtained from the data from Moors and colleagues do not reach significance. Although this may also have to do with the fact that they are collected at the level of the word form, rather than at the level of the concept and that they only cover 22.03% of the dialect data, visual inspection of the correlation between these valence measures and the response variable indicates that the positive and negative concepts behave similarly. However, as these ratings are organized along a continuous axis from very negative to very positive, determining where negative ends and neutral begins, and where neutral ends and positive begins, is relatively difficult and, to some extent, arbitrary. Second, the author manually coded the concepts of the six semantic fields on two separate occasions (once in 2014 and once in 2017). Preliminary analyses of the influence of these manual ratings on the amount of lexical heterogeneity in the WLD showed that the difference between negative and positive concepts concerning the amount of lexical geographical heterogeneity is not significant, whereas the difference between neutral and non-neutral concepts does

19 Participants in the arousal condition rated the degree to which a word refers to something active/arousing or passive/calm. The dominance ratings concern the degree to which a word refers to something weak/submissive or strong/dominant.

20 For this procedure, we first use the names for the concepts in the dictionaries to automatically match identical concepts to each other. For concepts that were not automatically linked this way, we manually verified whether they have a counterpart in the other dictionary by relying on the most frequent question from the NCDN questionnaires that was used to elicited the dialect names for the concept.

21 The concept names were presented in a more clear form if necessary. The concept GEVOELIG (ZIJN) '(to be) sensitive' was named GEVOELIG ZIJN 'to be sensitive' (without parentheses) and ETENSKETELTJE 'kettle for food' was called ETENSKETELTJE, OM ETEN NAAR ARBEIDERS IN HET VELD TE BRENGEN 'kettle for food, to bring food to the workers in the fields'.

22 These definitions are a based on the question from the NCDN questionnaires that was used most frequently per concept to elicit the dialect names. Each question was manually examined for accuracy and clarity, and modified if necessary.

have a large effect. Consequently, using a numerical scale in the rating task that ranges from very negative to very positive, like in the Moors and colleagues data, would cause the same problems for the analysis that occur for the valence data: the continuous variable would probably not show a linear relationship to the response variable, as both the low and high values of the variable impact the amount of variation a concept shows, but re-categorizing the continuous scale is problematic, because determining where negative ends and where positive begins, is relatively arbitrary. For this reason, we chose a more restricted set-up, in the form of a forced-choice task with only three categories (and an 'uncertain' value).

To calculate inter-rater reliability, we used Light's kappa (Hallgren 2012, Levshina 2015), an adjusted version of Cohen's kappa for more than two coders. This measure takes values from -1 to 1, with 1 indicating complete agreement, 0 completely random agreement and -1 complete disagreement. The measure reaches a value of 0.675, which indicates a moderate to substantial amount of agreement (Hallgren 2012: 6).[23] Disagreement between the raters mostly stems from their individual cut-off points of non-neutrality, rather than from contradictory choices regarding the positive or negative valence of a concept: some raters call a concept positive or negative more quickly than others.[24] For example, for the concept DIABOLO 'diabolo (a type of toy)', three raters indicate that the concept is neutral, while two raters code it as positive; the concept LIJKWAGEN 'a hearse' is rated as negative by four coders and as neutral by the fifth one. For only 59 concepts out of the 1935, contradictory choices between the positive and negative valence of the concept occur. Consequently, we operationalize affect as a measure of certainty regarding the non-neutrality of a particular concept. More specifically, it is calculated as the **proportion of non-neutral ratings** per concept. For instance, for the concept BARENSWEEËN 'labour pains', five rating scores are available, with four of these indicating that the concept is non-neutral (viz. negative), the value for affect-sensitivity is 0.8. If the ratings are categorized as neutral versus not-neutral (i.e. positive or negative) in this way, Light's kappa is equal to 0.645.

Additionally, to verify the validity of the ratings provided by our participants, we determine the degree to which they correlate with the 691 concepts for which mean valence information is available from the data collected on a much larger scale by Moors and colleagues. Figure 3.2 shows the relationship between the two variables. On the one hand, the green line on the figure, which is a regression line that represents the linear relationship between the two variables, indicates that a moderate negative correlation is found between the mean valence ratings (on the x-axis) and our own variable on the y-axis (Spearman's rho = -0.255; $p < 0.001$). More specifically, the participants in our rating experiment agree somewhat more about the non-neutrality of highly negative concepts, which have a low value on the x-axis, than about the highly positive concepts, with higher values on the x-axis. However, the overall validity of the proportion of non-neutral ratings is substantiated by the loess smooth (indicated with the red non-linear line and the corresponding dashed red confidence bands), a non-parametric regression method that allows for non-linearity between the two variables. The U-shape of this smooth clearly confirms that both the positive and negative concepts that occur in the data from Moors and colleagues have a much higher value for our own variable, proportion of non-neutral ratings, than the neutral concepts towards the centre of the x-axis.

The proportion of non-neutral ratings per concept ranges from 0 to 1, with mean 0.427 and standard deviation 0.427. Surprisingly, as a single rating value was elicited per participant for concepts that occur in both dictionaries, it differs significantly between the WLD and WBD ($t = 2.6087$, $df = 3077.7$, $p < 0.01$). In the WLD, the mean is 0.448, with standard deviation 0.424; the mean in the WBD is 0.408, with standard deviation 0.429. However, this finding is related to the fact that concepts from the field of family & sexuality are unavailable for the WBD. If these concepts are excluded, the difference is no longer significant ($t = 1.736$, $df = 2851.8$, $p > 0.05$). Concepts with a value of 1 for proportion of non-neutral ratings include LIEFKOZEN 'to caress', OPSCHEPPERIJ 'bragging' and OUD, BOUWVALLIG OF ARMOEDIG HUIS 'old, shabby house'. A value of 0 is found for concepts like KINDERSTOEL 'baby chair', MIAUWEN 'to miaow' and MIDDELMATIGE DUIF 'mediocre pigeon (in pigeon keeping)'.

23 Light's kappa was calculated on the full dataset by only using the ratings from the three coders who completed ratings for all six semantic fields.

24 Additionally, we calculated Light's Kappa per semantic field to also compare the validity of the measure for the raters who only coded a subset of the semantic fields. The analysis indicates that Light's Kappa differs between semantic fields: it takes values between 0.326 and 0.746. Low kappa values are found for the semantic fields of celebration & entertainment (0.326) and family & sexuality (0.349), two fields that contain a relatively large amount of obsolete concepts (e.g. AARDAPPELFOOI (festivities held after the digging up of the potatoes), BRUIDSJONKER 'page (at a wedding)', DOOPMUTSJE 'bonnet worn during a baptism', HOGE HOED BIJ BEGRAFENIS 'top hat worn at a funeral' and highly specialized concepts relating to pigeon keeping, while high values occur for fields with more universal or modern concepts (personality & feelings: 0.746, society, school & education: 0.667). Furthermore, the differences are again predominantly related to the raters' individual cut-off points of non-neutrality, especially for concepts that are less well known in modern-day life.

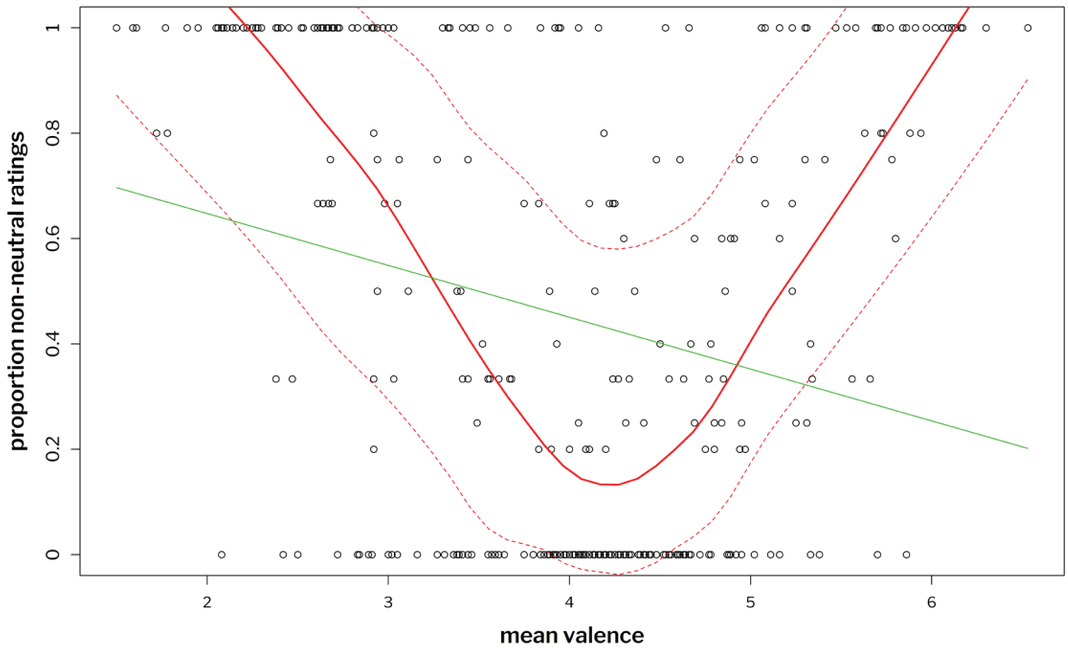


FIGURE 3.2
Correlation between mean valence (Moors et al. 2013) and proportion of non-neutral ratings

3.3 DEPENDENT VARIABLE

The dependent variable used in the analysis is operationalized in the same way as in the pilot studies to ensure that the results are comparable. This variable, lexical geographical heterogeneity, is a composite variable that takes into account several aspects of the degree of lexical diversity in the dialect lexicon. The calculation of the variable consists of several steps.

First, the variable considers **the number of unique lexemes** (i.e. types) per concept.[25] This variable ranges from 1 to 202 with mean 23.3 and standard deviation 20.588. Concepts with a high value for this variable are RIJK ZIJN 'to be rich' (WLD: 202 unique lexemes; WBD: 191 unique lexemes) and AARZELEN, TREUZELAAR 'to hesitate, dawdler' in the WBD (194 unique types).[26] Only one unique lexical item is used for concepts like BLOED 'blood' in both dictionaries, ORGEL 'pipe organ' in the WLD and TEEN 'toe' in the WBD. Alternative operationalizations of this measure can be envisaged that take into account the number of tokens (i.e. responses) per concept and the frequency with which each unique lexical item occurs (for examples, see Cornips et al. 2016, Geeraerts, Grondelaers & Speelman 1999 and Swanenberg 2004, 2010). However, as we assume that the number of available responses to some extent reflects the degree of salience of a concept (see the variable 'proportion of missing places' above), and because we aim to keep the analysis comparable to the pilot studies, we rely on the calculation used in these studies. However, alternative operationalizations will be examined in chapters 5 and 6.

Second, the operationalization of the response variable also reflects the fact that dialectal data are geographically stratified. In dialectometry, a subfield of dialectology, a number of ways to objectively quantify the linguistic distance between different dialects of a single overarching language have been devised (a fruitful line of research in typology uses comparable methods to measure the distance between less closely related varieties, e.g. Brown et al. 2009, Cysouw & Comrie 2009, Jäger 2013). Geographical scatter is traditionally operationalized in dialectometry as the linguistic distance between pairs of locations. Linguistic distance has been measured in several ways and in different types of datasets, including a binary operationalization of (dis)agreement between locations (Séguy 1971), relative or weighted values of identity (Goebl 1984, 2010), Levenshtein distance (e.g. Heeringa 2004, Wieling, Nerbonne & Baayen 2011), spatial autocorrelation (Grieve, Speelman & Geeraerts 2011), and Euclidian distance (e.g. Szmrecsanyi 2008). Additionally, a second line of dialectological research considers perceptual beliefs of laymen about linguistic distance (Preston 1999,

25 In the pilot studies, this aspect of the response variable is called 'lexical diversity'. However, in this dissertation, we reserve this term for the amount of lexical variation a concept shows in the dialect lexicon in general. We rely on more specific terms, like the number of unique lexemes or types per concept, to refer to the number of unique variants that occur for a concept.

26 In the Limburgish data, AARZELEN and TREUZELAAR are included as separate concepts. AARZELEN occurs with 36 types in this dataset, while for TREUZELAAR 73 unique lexical items are recorded.

Weijnen 1946). More recently, scholars have inquired into the degree to which variants are scattered in a more or less heterogeneous way across geographical space, by taking into account the relationship between linguistic and geographical distance directly. Rumpf et al. (2010), for instance, have devised a method to automatically detect similar spatial patterns between the distribution of lexical variants for different concepts in a dialectological atlas. Cornips et al. (2016) include a measure of entropy to model the degree of geographical scatter of a particular concept. The aim of most of the studies outlined here (the research of Cornips and colleagues forms the exception), is to detect dialect regions within a larger dialect (or language) variety. The measure used in the pilot studies can be considered as belonging to the third type, because it also directly models geographical fragmentation, i.e. the relationship between linguistic and lexical distance. However, instead of using it to distinguish dialect areas or isoglosses in the Limburgish and Brabantic dialect area, we examine features that influence variation in **geographical fragmentation** (but see Speelman & Geeraerts 2008 for the advantages of including concept-based features for the purpose of dialectometric research).

In the pilot studies and in this and the next chapter, the operationalization of the degree of geographical fragmentation a concept shows, takes into account the average geographical distance between two locations with the same variant (**dispersion**) and the average geographical surface of a particular lexical item (**range**) per concept. Conceptually, the degree to which a particular concept shows dispersion concerns the degree to which, on average, the distribution of the lexical variants for the concept is characterized by the interference of other lexemes that are used for the same concept. A concept is highly dispersed if the lexical variants are scattered across geographical space in a heterogeneous way, without the formation of clear areas where a particular variant is used consistently, but rather with several variants used intermittently (en example can be found in Figure 3.3). Little dispersion occurs if homogeneous areas can be distinguished, as in Figure 3.4. Dispersion per concept is calculated by relying on the average degree of dispersion of all the lexical items for the concept. First, for every location where a particular variant occurs, the distance to the nearest location (in terms of geographical distances) with the same variant for the same concept, is measured. Then, the average of these distance to the nearest location with the same word is calculated for all the locations where the lexical item under scrutiny occurs. Second, again for every location where the variant is found, the distance to the nearest location with any observation for the concept is taken into account. Again, the average of this measure is calculated as well. The dispersion of the lexical variant is then quantified by dividing the first averaged figure (average distance to the nearest location with the same term) by the second one (average distance to the nearest location with any observation for the concept). The dispersion of a concept is subsequently calculated as the average amount of dispersion for all the lexical items that occur for the concept. This average is also weighted by the relative contribution of each lexical item to the onomasiological profile of the concept.[27] Dispersion ranges from 1 to 4.401, with mean 1.894 and standard deviation 0.590.[28] Concepts with a high degree of dispersion include VERBEUZELEN 'to squander' in the WLD (4.401) and IEMAND WEERSTAAN 'to resist someone' in the WBD (4.323; Figure 3.3). Concepts like BLOED 'blood' in both dictionaries, JOJO 'yo-yo' in the WLD and GETUIGE 'witness' in the WBD, have a value of 1 for dispersion. Figure 3.4 shows the distribution of the variants for the concept SLUIS in the WBD, which also has a low value for dispersion (1.120).

Next to dispersion, geographical fragmentation is also influenced by the range of a concept, because concepts for which the average range of the lexical variants is high, like in Figure 3.4, are less heterogeneous than concepts with a lower geographical range. An example of a concept of the latter type is provided in Figure 3.5. Only lexical items that occur more than three times for this concept are plotted with a coloured triangular symbol. Other locations where data for the concept occurs, are indicated with a grey circle.

To calculate the geographical range of a concept, we rely on the relationship between, on the one hand, the average geographical area spanned by the lexical items for the concept and, on the other hand, the total area where the concept occurs. Most concepts in the dataset occur in the entire Limburgish or Brabantic dialect area. Only about 10% of the concepts span less than 75% of the surface of one of these dialect regions and only 37 (out of the 3136) concepts occur in less than 40% of these areas. The latter group of concepts all belong to the Brabantic dialect data. Furthermore, most of them have a more limited geographical scope because of clear reasons. First, data for 21 out of the 37 concepts, all belonging to the semantic field 'the house', were collected with NCDN questionnaire N 104 (2000). However, this questionnaire was

27 As explained in chapter 1, an onomasiological profile takes into account the relative frequency of each lexical item used for a particular concept (Speelman et al. 2003). For instance, if for a particular concept Z, three lexical items occur with differing frequencies (a: 10 observations, b: 60 observations and c: 30 observations), the relative contribution of each variant is calculated as the number of observations per variant divided by the total number of observations for the concept. For variant a, the relative contribution is, thus, 0.10; for b, it is 0.6 and for c it is equal to 0.3.

28 The geographical distances and areas calculated in this chapter are not expressed in (squared) kilometres. We use a different coordinate reference system available in the geolocation data of the WBD and WLD than longitude and latitude.

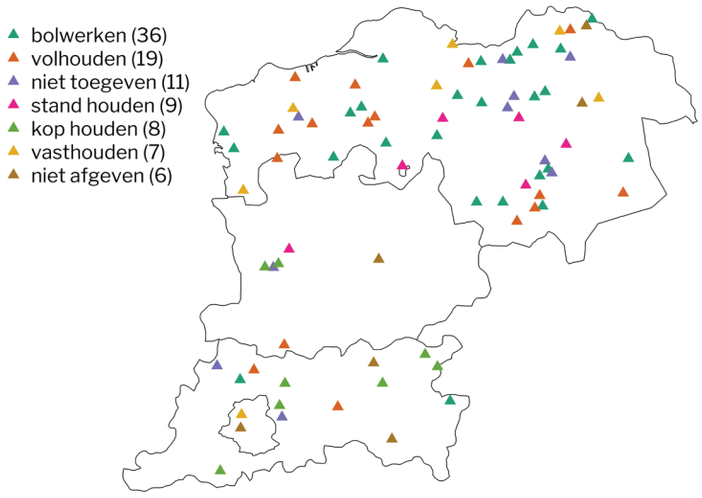


FIGURE 3.3
Geographical distribution of the lexical variants for IEMAND WEERSTAAN *in Brabant (only items that occur more than 5 times in the WBD are included)*

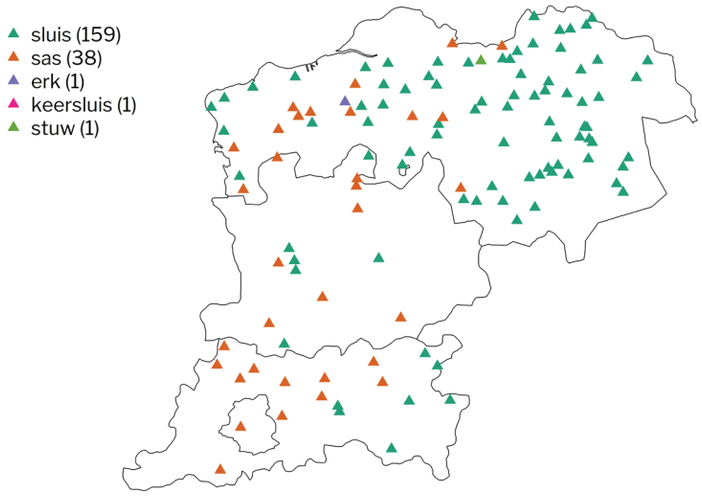


FIGURE 3.4
Geographical distribution of the lexical variants for SLUIS *in Brabant*

predominantly distributed in the province of North Brabant in the Netherlands for the WBD. Consequently, for these concepts, data from Belgium are not available. Second, some concepts only occur in a limited geographical region, due to the nature of the referent to which they refer. More specifically, 12 out of the 37 concepts are related to the Belgian monetary unit that was used before the introduction of the euro (e.g. 1 BEF (ZILVER) '1 silver Belgian frank', 100 BEF (BILJET) 'note of 100 Belgian Franks' and NAPOLEON 'gold coin of 20 Belgian franks'). Data for these concepts were only elicited in the northern part of Belgium for the WBD (the differences are less outspoken in the WLD). Differences in the geographical span of the concepts are taken into account in the operationalization of the range of a concept.

In practice, the range of a concept is calculated as follows. First, for each lexical item for a particular concept, the area where the word occurs, is measured to obtain the range per lexical item. This area is subsequently divided, per lexical item, by the total area where the concept occurs. The geographical range for a specific concept consists of the average of the range for all the lexemes used for the concept. However, for this variable, the importance of each lexical item for the aggregated variable is weighted in

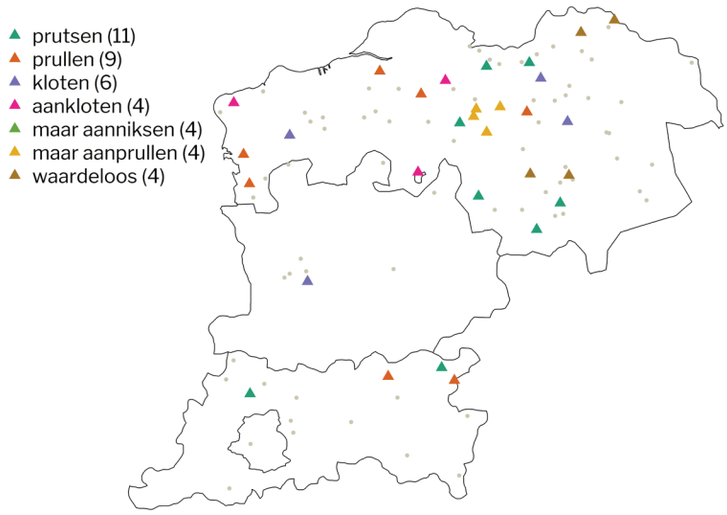


FIGURE 3.5
Geographical distribution of the lexical variants for NUTTELOZE ARBEID VERRICHTEN; NUTTELOOS WERK. *Only items that occur more than 3 times in the WBD are plotted. To show the total range of the concept (85.29%), other locations for which data is available, are indicated with a grey circle*

function of its relative contribution to the concept profile as well. Geographical range per concept takes values between 0.003 and 1 in the dataset, with mean 0.579 and standard deviation 0.229. Concepts with a low value for range include VERSCHILLENDE KNIKKERSPELEN in the WLD (0.003), a concept for which respondents were asked to provide names for different games of marbles that they are familiar with, and NUTTELOZE ARBEID VERRICHTEN; NUTTELOOS WERK 'to mess about' in the WBD (0.076; Figure 3.5). A value of 1 for range is found for concepts like LUISTEREN 'to listen' in the WLD and GETUIGE 'witness' in the WBD. An example of a concept with a value for range that is only slightly lower than 1, is found in Figure 3.4: SLUIS has both a low value for dispersion and a relatively high value for weighted averaged range (0.879).

The composite variable, lexical geographical heterogeneity, is calculated by means of the following formula:

$$\textit{lexical geographical heterogeneity} = \textit{number of unique types} * \frac{\textit{dispersion}}{\textit{range}}$$

Dispersion is divided by range, because range can be considered as the inverse of the lack of spread of a concept. As all three aspects, number of unique types, dispersion and 1/range per concept, contribute to the degree of lexical heterogeneity, the product of these measures is used. In practice, we use the natural logarithm of lexical geographical heterogeneity, because this results in a more linear relationship between the predictor variables and the response. The logarithm of lexical geographical heterogeneity ranges from 0 to 9.560 with mean 4.054 and standard deviation 1.507. Concepts with a high value for the response variable are VERSCHILLENDE KNIKKERSPELEN 'various games of marbles' (9.513) and ZWAK EN MAGER PERSOON 'weak and meagre person' (9.560) in the WLD and GELUIDLOOS EEN WIND LATEN 'to let off a fart silently' (8.652) and OP DE ZENUWEN WERKEN 'to enervate' (8.562) in the WBD. Concepts with a value equal to zero (i.e. concepts with one lexical item that occurs everywhere) include BLOED 'blood' in both dictionaries, ADER 'vein' in the WLD and EED 'oath' in the WBD.

3.4 METHODOLOGY

The impact of the predictors on the response variable is assessed using linear regression analysis. To construct the model, we rely on an automatic backward selection algorithm, that uses the AIC criterion of the models and that allows for two-way interactions between all the predictors in the model. We subsequently exclude model terms and interactions that do not significantly contribute to the explanatory power of the model. Additionally, we remove interactions between single predictors that model the same overarching variable to avoid a harmful amount of multicollinearity in the model. For instance, we do not allow for an interaction between proportion of multi-word expressions per concept and proportion of hapaxes per concept, as both of these variables model the influence of lack of onomasiological salience. Finally, we avoid overfitting by only retaining the interactions that contribute enough to the reduction of the variation in the response variable.

Before interpreting the results of the model, the assumptions were examined. More specifically, we verified that the relationship between the predictors and the response is linear, that there are no outliers, that the model does not suffer from heteroscedacity, that there is no harmful multicollinearity between the predictors and that the residuals are normally distributed. The model suffers from some heteroscedacity (due to the fact that the data only contain a few observations with a very high value for the response variable) and the relationship between the response and some explanatory variables is only linear-like, but as the results are stable under bootstrapping, using the procedure outlined in Levshina (2015: 167-169), the predictors in the model are robust.

3.5 RESULTS

The output of the regression model is shown in Table 3.4. All the fixed-effects predictors that were discussed above, reach significance. The adjusted R^2 value of the model is 0.7311, which indicates that about 73% of the variance in the response variable is explained by the combination of the predictors and interaction effects.

To interpret the partial contribution of each predictor to the explanatory power of the model, we start with the variables that are not included in any interaction effect. The results for the variables that are included in an interaction, will be inspected visually below. Overall, the model confirms the results obtained in the pilot studies: in a different dialect area and in other semantic fields, more variation is found for onomasiologically more vague and less salient concepts and for concepts that are prone to affect.

3.5.1 Main effects

The first variable that is not included in an interaction effect is the dictionary to which the concept belongs. The reference level for this variable, WLD, is included in the intercept. To determine the effect of the variable, we compare the estimate (in the second column of the table) and the p-value (in the last column) for the alternative level, WBD, to the intercept. The fact that the p-value for the alternative level is smaller than the alpha level, 0.05, indicates that the amount of lexical geographical heterogeneity differs significantly between the dictionaries, if all the other variables are taken into account. The direction of this effect is reflected by the estimate (0.184). The positive sign reveals that significantly more variation is found in the Brabantic data. Furthermore, the absolute value of the estimate demonstrates the effect size for the predictor: if all the other variables are stable, lexical geographical heterogeneity takes a value 0.184 higher in the WBD than in the WLD. As it is unlikely that, for concepts with similar properties, dialect speakers from Brabant use more different words than people from Limburg, the divergence between the dictionaries is probably related to other differences, like the fact that the Brabantic dialect area is larger and that more data from the WBD are available. Crucially, however, because no significant interaction effects are found with any of the other predictor variables and 'dictionary', the model confirms that the effect of the semantic concept features is stable across the two datasets.

The second set of variables in the table that are not included in an interaction effect, viz. proportion of missing places, proportion of MWE's and the binary operationalization of prevalence, concern the lack of salience of a concept. The small p-value and the negative sign of the estimate (-1.055) for the former predictor indicate that a significant negative correlation is found between the proportion of missing places for a particular concept and the amount of variation the concept shows. Although this is not the effect that we expected, as we assumed that concepts with a higher amount of missing places are less salient and, thus, predicted a positive correlation, the effect of this variable was the same in the pilot studies. In these studies, this finding was explained by the fact that a larger amount of missing places may also result in a smaller amount of lexical variation per concept, due to the fact that less data are available.

Furthermore, for the other variables that measure the lack of salience of a concept, we do find the expected effect: less salient concepts show significantly more variation across the dictionaries and in every semantic field. First, the model finds a significant positive correlation between the proportion of MWE's per concept and the amount of lexical geographical variation, as indicated by the positive estimate (0.583) and the small p-value. Second, the results also indicate that significantly more variation is found for non-prevalent concepts or concepts that are not available in the prevalence data collected by Keuleers et al. (2015), in comparison to the concepts that are prevalent (the reference level, included in the intercept).

The last variable that is not included in any interactions, proportion of non-neutral ratings, measures the degree of certainty concerning the affect-sensitivity of a concept. The higher the proportion of non-neutral ratings for a particular concept, the more certain we can be that the concept is prone to affect. The estimate (0.280) and small p-value for this variable indicate that the effect is as expected across dictionaries and semantic fields: significantly more variation is found for concepts that are prone to affect.

model term	estimate	SE	p-value
intercept	2.586	0.072	< 0.001
dictionary			
WBD	0.184	0.032	< 0.001
semantic field			
the house	0.344	0.082	< 0.001
celebration & entertainment	0.059	0.079	NS
personality & feelings	0.200	0.090	< 0.05
family & sexuality	0.132	0.121	NS
society, school & education	0.274	0.072	< 0.001
lack of salience			
proportion of missing places	-1.055	0.104	< 0.001
proportion of MWE's	0.583	0.076	< 0.001
proportion of hapaxes	13.318	0.552	< 0.001
prevalence binary (*missing / not prevalent*)	0.228	0.032	< 0.001
vagueness			
lexical non-uniqueness	0.032	0.003	< 0.001
affect			
proportion of non-neutral ratings	0.280	0.042	< 0.001
interaction terms			
sem. field (*the house*) : proportion of hapaxes	1.483	0.792	< 0.1
sem. field (*celebration & entertainment*) : proportion of hapaxes	-3.220	0.638	< 0.001
sem. field (*personality & feelings*): proportion of hapaxes	-1.867	0.626	< 0.01
sem. field (*family & sexuality*) : proportion of hapaxes	0.736	1.205	NS
sem. field (*society, school & education*) : proportion of hapaxes	-1.195	0.639	< 0.1
sem. field (*the house*) : lexical non-uniqueness	-0.002	0.004	NS
sem. field (*celebration & entertainment*) : lexical non-uniqueness	0.018	0.006	< 0.01
sem. field (*personality & feelings*): lexical non-uniqueness	-0.012	0.003	< 0.001
sem. field (*family & sexuality*) : lexical non-uniqueness	-0.007	0.010	NS
sem. field (*society, school & education*) : lexical non-uniqueness	0.001	0.003	NS
proportion of hapaxes : lexical non-uniqueness	-0.065	0.007	< 0.001

TABLE 3.4
Output of the regression model

3.5.2 Interaction effects

Next, we turn to the first interaction effect included in the model, between semantic field (reference level 'the human body') and proportion of hapaxes (an operationalization of lack of salience). A visualization of the effect of this interaction is presented in Figure 3.6. The figure shows the predicted effect of proportion of hapaxes (in different colours) on the response variable, lexical geographical heterogeneity, on the y-axis, per semantic field (on the x-axis). Although the interaction effect is significant, proportion of hapaxes has the same effect in every semantic field: more variation is found for concepts with a high value for this predictor (i.e. concepts that are less salient). However, the degree to which higher values of proportion of hapaxes affect the response variable, differs between the semantic fields. More specifically, in comparison to the reference level, the proportion of hapaxes per concept affects the other universal field (personality & feelings) and the socially-bound semantic fields (celebration & entertainment and society, school & education) significantly less. The impact of the predictor is significantly higher in the locally-bound semantic field 'the house'.

The interaction effect corroborates the findings discussed in Pickl (2013), that fields that are not locally bound show more lexical levelling (i.e. less lexical geographical variation) than fields containing concepts that are predominantly relevant in a local community (recall that a similar interpretation may be relevant for the distribution of the variant on the *house/mouse* map of Kloeke, see chapter 1). Furthermore, the difference between the locally and non-locally bound semantic fields is larger for concepts that are less salient. An explanation offered by Pickl is that locally-bound concepts are discussed less often on a large socio-geographical scale and that the speaker who would typically discuss the concept, is not very mobile. The interaction effect present in our model, however, shows that aspects of the prototype-theoretical structure of the lexicon can enhance this effect.

To exemplify this interpretation, Table 3.5 contains the concepts with the highest proportion of hapax legomena per semantic field. The table confirms that these concepts are not very salient (although a few of them probably also have a high proportion of hapaxes for euphemistic reasons, like BOEZEM 'bosom' and GESLACHTSGEMEENSCHAP HEBBEN 'to have sexual intercourse'). Additionally, the concepts belonging to the locally-bound semantic fields, the house and family & sexuality, are concepts that are probably not discussed often in informal conversations across town borders or on a supra-local scale, while the concepts belonging to the other semantic fields are more socially relevant.

The second interaction effect included in the model concerns the influence of lexical non-uniqueness depending on the semantic field to which a particular concept belongs (Figure 3.7). As explained above, lexical non-uniqueness measures how often a lexical item that occurs for the concept under scrutiny is used to refer to other concepts as well. Again, the effect of lexical non-uniqueness is the same across semantic fields: vaguer concepts show more lexical geographical heterogeneity. However, the figure indicates that differences in the effect size of the variable exist between semantic fields.

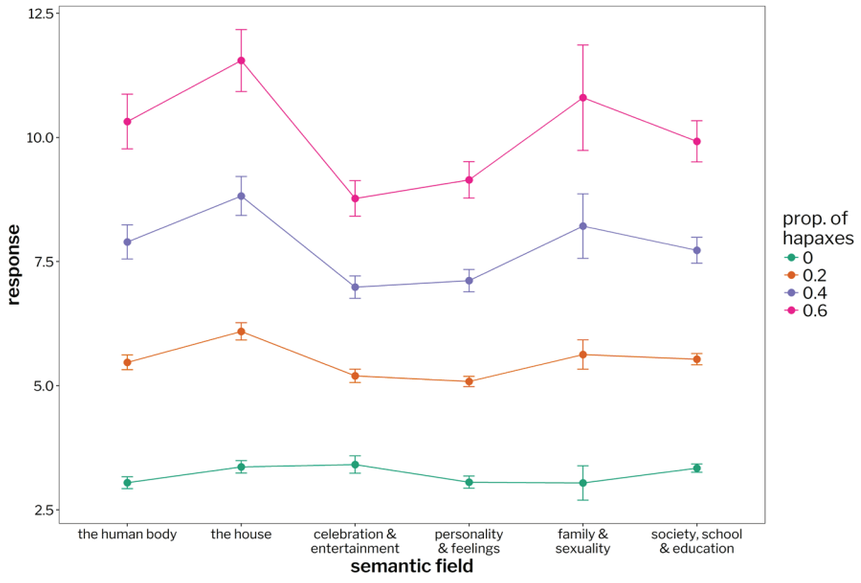


FIGURE 3.6
Interaction between semantic field and proportion of hapaxes

the human body	the house	celebration & entertainment
BOEZEM 'bosom'	**KROLLEN** 'to caterwaul'	**VERSCHILLENDE KNIKKERSPELEN** 'various games of marbles'
GEDRONGEN PERSOON 'stocky figure'	**BOT MES** 'blunt knife'	**KNIKKEREN** 'to play marbles'
ZWAK EN MAGER PERSON 'weak and meagre person'	**HALFROND PLANKJE MET STEEL WAARMEE DE AS IN DE ASBAK WERD GETROKKEN** 'piece of wood to get the ashes into the ashpan'	**STAARTDUIF** 'pigeon listed as one of the last winners (pigeon keeping)'
(MET) STEVIGE BENEN '(with) hefty legs'	**HOUTEN SPAANTJES WAARMEE MEN VUUR NEEMT UIT DE KACHEL** 'chips of wood to get fire from the stove'	**VAN ACHTEREN KOMEN** 'pigeons coming from the opposite direction (pigeon keeping)'
GELUIDLOOS EEN WIND LATEN 'to let off a fart silently'	**LICHT ONTVLAMBAAR MATERIAAL IN DE TONDELDOOS** 'inflammable material in the tinderbox'	**VAN DE VERKEERDE KANT KOMEN** 'pigeons coming from the wrong direction (pigeon keeping)'
personality & feelings	**family & sexuality**	**society, school & education**
GEMAKKELIJKSTE WIJZE, GEMAKKELIJKST, GEMAKKELIJK MAKEN 'easiest (way), to make something easy	**GESLACHTSGEMEENSCHAP HEBBEN** 'to have sexual intercourse'	**WELBESPRAAKT ZIJN** 'to be eloquent'
ZICH HEEL WAT INBEELDEN, INGEBEELD PERSON 'to fancy oneself'	**KIND (TROETELNAAM)** 'child (kinship terms)'	**BEKAKT PRATEN** 'to talk posh'
BETROUWBAAR IEMAND 'reliable person'	**MANZIEK** 'nymphomaniacal'	**BRASSEN** 'to binge'
ONHEILSPELLEND; SLECHT NIEUWS 'ominous, bad news'	**PUBER** 'adolescent'	**WINKEL DRIJVEN** 'to run a shop'
OP DE ZENUWEN WERKEN 'to enervate'	**VROUWZIEK** 'crazy about women'	**ZICH AANSTELLEN** 'to show off'

TABLE 3.5
Concepts with the highest value for proportion of hapaxes (lack of salience) per semantic field

More specifically, the impact of lexical non-uniqueness seems to be greater for the three semantic fields with a large amount of concrete concepts on the left of the figure (the human body, the house and celebration & entertainment) than on the three more abstract semantic fields to the right of the figure (personality and feelings, family & sexuality and society, school & education). This may have to do with the fact that, for concrete concepts, perceptual information is available which may make them relatively easily distinguishable from related concepts (at least towards "co-hyponymous" concepts, at the same taxonomical level). By contrast, for more abstract concepts, such perceptual clues are not available, which may result in these concepts being more vague towards related items on the same taxonomical level in general. Consequently, if abstract concepts are always more vague than concrete ones, this reduces the effect that lexical non-uniqueness *can* have on the amount of variation these concepts show. However, at this point, the degree of concreteness is measured per semantic field and by relying on concreteness ratings for words. This explanation can only be corroborated further by means of concreteness ratings per concept.

Additionally, these results need to be attenuated. First, the error bars indicate that results are less reliable in every semantic field for concepts with a very high value for lexical non-uniqueness. However, the difference between the concrete and abstract semantic field remains stable and significant in a model that only includes concepts with a value lower than 50. Additionally, preliminary analyses indicated that significant differences in the amount of

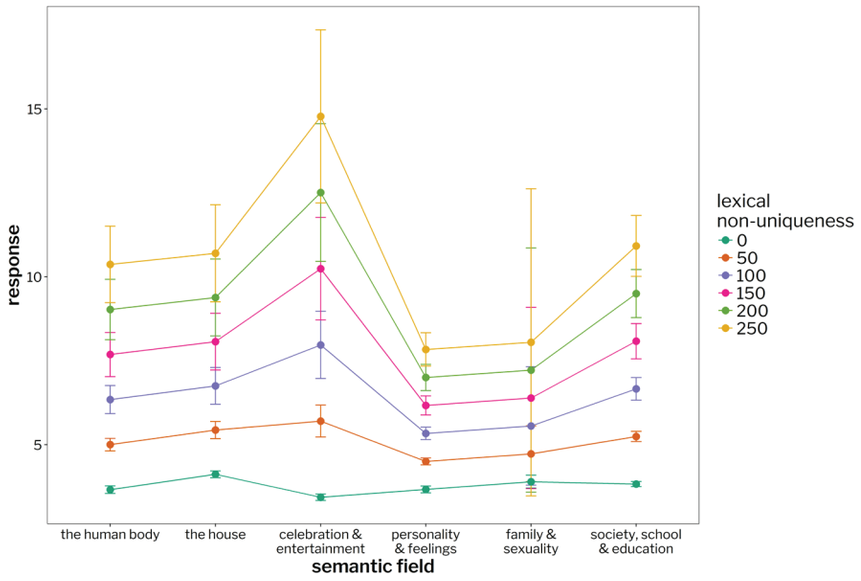


FIGURE 3.7
Interaction between semantic field and lexical non-uniqueness

lexical non-uniqueness occur between the semantic fields: only a few concepts have a relatively high value for lexical non-uniqueness in the semantic fields of celebration & entertainment (min.: 0, max.: 40, mean: 5.686, sd: 7.376) and family & sexuality (min.: 0, max.: 35, mean: 5.681, sd: 8.465). Consequently, the predicted values that are shown in Figure 3.7 for these semantic fields, are only based on observations for a limited set of concepts (as indicated by the error bars). The concepts that are the most vague in the field of celebration & entertainment all relate to the same topic, viz. children's games, especially playing with marbles (e.g. STUIKEN 'to throw marbles in a pit', KNIKKER 'marble', KLEINE KNIKKER 'small marble', BONKEN 'to let marbles bounce against each other'). It is not surprising that these concepts show a high amount of lexical geographical heterogeneity as children's games are known to be lexically diverse due to the creative nature of child language. A similar pattern shows up for the semantic field of family & sexuality: the concepts with the highest value for lexical non-uniqueness are also limited to specific semantic subfields. They either concern a young person (KIND (TROETELNAAM) 'kinship terms for a child', PUBER 'adolescent', MEISJE MET WIE EEN JONGEN VERKERING HEEFT 'girlfriend of a boy'), or indecency (ONKUIS 'indecent', MANZIEK 'nymphomaniacal').

The final interaction effect in the model concerns the relationship between lexical non-uniqueness (a measure of vagueness) and proportion of hapaxes (a measure of lack of salience). Figure 3.8 shows the effect of this interaction. On the x-axis, lexical non-uniqueness per concept is provided. The influence of proportion of hapaxes per concept on the response variable (y-axis) is indicated with coloured lines (and corresponding error bars and confidence bands). The figure shows that the influence of lexical non-uniqueness is especially large for concepts that are highly salient (i.e. a low proportion of hapaxes). If these concepts are relatively vague, they show much more variation than their non-vague counterparts. However, the impact of lexical non-uniqueness decreases for less salient concepts. Because the number of concepts with very high values for lexical non-uniqueness is relatively small, the confidence bands towards the right of the figure overlap. However, the effect remains stable in a model that only includes concepts with a value smaller than 50 for lexical non-uniqueness. Additionally, most of the concepts have a relatively low proportion of hapaxes: only 184 out of the 3136 concepts in the database have a proportion of hapaxes larger than 0.25. Consequently, for the highly non-salient concepts, the effect is only based on a small number of observations.

Perhaps this finding can be explained by the fact that non-salient concept show a lot of variation overall, regardless of whether or not they are vague. For instance, for the non-salient (proportion of hapaxes = 0.567), but also non-vague (lexical non-uniqueness = 4) concept STAARTDUIF 'pigeon listed as one of the last winners (pigeon keeping)' in the WBD, a lot of variation still occurs. For instance, 41 lexical items exist for the concept, which show a lot of geographical fragmentation (dispersion = 2.09, range = 0.13). Out of these 41 lexemes, 37 do not occur for other concepts in the

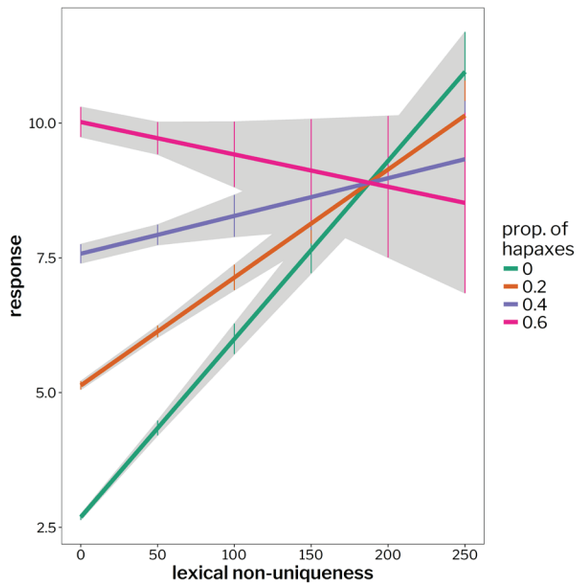


FIGURE 3.8
Interaction between lexical non-uniqueness and proportion of hapaxes

WBD. This indicates that many of these lexical items are used hesitantly, which results in the smaller impact of lexical non-uniqueness for less salient concepts.

For more salient concepts, however, vagueness does play a larger role: the difference between non-vague and vague concepts is much larger. For non-vague salient concepts, there is not a lot of variation in general: one or a few lexical items take up a very strong position in the profile for the concept and not many other lexical items occur. This is, for instance, the case for BLOED (0 hapaxes, 0 lexical items that occur for other concepts as well), for which 1 variant, *bloed*, is used everywhere. For vaguer salient concepts, however, additional variants do occur. For example, for the salient and vague concept SLIM (ZIJN) '(to be) smart', one lexical item is relatively frequent throughout the Brabantic dialect area, viz. *slim*: it is used in 48.97% of all observations for the concept. Additionally, however, 69 other lexical items exist that all contribute much less to the geographical profile of the concept.

3.6 DISCUSSION

The aim of this chapter was to provide further evidence for the findings of the pilot studies, that lexical diversity is not only influenced by lectal conditions, but that the prototype-theoretical structure of the lexicon influences the amount of variability a concept shows as well. The results outlined in this chapter confirm this view. Concepts with a higher degree of onomasiological vagueness and lack of salience show more lexical geographical variability. Furthermore, the degree to which a concept is prone to affect, which can be considered an aspect of the encyclopaedic nature of meaning, influences lexical geographical heterogeneity as well. Crucially, overall, the number of interaction effects is limited, which indicates that the influence of onomasiological salience, vagueness and proneness to affect is very strong. This is further confirmed by the fact that the regression model indicates that these concept features are stable in another dialect area than the region on which the pilot studies focused, as we find no significant interactions between any of the concept features and the dictionary to which the concepts belong. Additionally, although the model indicates that differences between semantic fields occur, the effect of the concept features remains the same across all six semantic fields. The interaction effects show that the magnitude of the influence of two specific predictors differs between these fields. First, socio-cultural conditions have an effect: locally-bound concepts show more lexical geographical heterogeneity, especially when they are not salient. The second interaction implies that the importance of semantic features between semantic fields is related to the processing of lexical material: the effect of lexical-non uniqueness is stronger for fields with, on average, a higher amount of concrete concepts. We explained this finding by referring to the fact that, perhaps, abstract concepts are more vague in general due to their language-based nature, which results in a reduction of the potential effect of lexical non-uniqueness. However, further research is necessary to corroborate this interpretation. Finally, we also found a significant interaction between two of the concept features, viz. proportion of hapaxes (lack of onomasiological salience) and number of non-unique types (onomasiological vagueness). We interpreted this finding by asserting that vagueness may predominantly affect salient concepts. Non-salient concepts (like STAARTDUIF) show a lot of variation in general, regardless of whether they are vague or not, which reduces the possible effect of lexical non-uniqueness on these concepts. However, non-vague onomasiologically salient concepts (like BLOED) generally do not show a lot of variation. When they are vague (e.g. SLIM (ZIJN)), the amount of lexical geographical heterogeneity increases dramatically.

Aside from supplementing the dataset with concept-based concreteness ratings, some further shortcomings should be addressed in follow-up research. First, in our operationalization of lexical non-uniqueness, we did not control for the polysemous use of lexical items. Although we calculated this variable per semantic field and per dictionary to reduce the chance of the polysemy, this method does not completely eliminate polysemy in the dataset. For instance, the lexical item *stop* occurs for two concepts in

the semantic field of the house in the WBD. It is used 107 times for the concept STOP 'plug for a bathtub or washbasin' and 73 times for the concept ZEKERING 'fuse'. It also occurs in Standard Dutch in both of these meanings. While these readings of the word *stop* are metaphorically related (both objects close of another object, viz. a bathtub and an electrical cord), they can be argued to be distinct enough for the item to be considered polysemous rather than vague. To disentangle polysemic readings from vague ones for the lexical items in the database, two methods can be envisaged. On the one hand, several logical, linguistic and definitional tests have been proposed to distinguish vagueness from polysemy. However, these tests have been shown to be problematic, because they can produce contradictory results and can be unstable between different contexts (Geeraerts 1993, 2015). An alternative, more objective approach would, therefore, be to consider the formulation of the questions used in the questionnaires, which are available in the dataset, as an operationalization of the definition of each lexical item. On the basis of these formulations, a large-scale quantitative method could be devised that uses the similarity between the words used in the questions to automatically estimate, for each lexical item in the database, the probability that it is vague (i.e. the formulation between questions is highly similar) rather than polysemous (i.e. the formulation between questions is highly distinct).

Other adaptations to the predictor variables may be envisaged. First, the results of both this chapter and of the pilot studies show that the proportion of missing places cannot be considered as a valid operationalization of the degree to which a concept is not salient. This variable always has the opposite effect on the response variable than the other measures of onomasiological salience. However, other measures can be used to gauge the degree to which a concept is salient. On the one hand, we can infer aspects of the environment of the dialect users by relying on non-linguistic, external data. This method will be used in chapter 6. On the other hand, we could rely on linguistic frequency data. However, as the data in the WBD and WLD represent the dialect lexicon of language users from the early 20th century and contain a relatively large amount of concepts that have, meanwhile, become obsolete or, at the least, less frequent, a large, highly diversified and historical dataset should be used for this purpose. Perhaps, the texts in *Nederlab*[29], a recently launched digital repository aimed at helping researchers examine changes in the language and culture of the Dutch language area, could serve as a reliable source for the types of concepts that are available in the dialectal data.

Furthermore, as the goal of this chapter is to model the spatial distribution of linguistic variants, this study bears some resemblance to research in ecology and spatial epidemiology, like disease mapping. The goal of this kind of research is generally to describe variation in the geographical spread of a particular disease, with geographical location being considered as an estimate of social, environmental and genetic risk factors (Elliot & Wartenberg 2001). A classical example, from Snow (1855, in Eyler 2013), concerns the local outbreak of cholera in the 1840s in London, which was uncovered to be caused by drinking water from a contaminated well. Recent studies in disease mapping rely on sophisticated methods to describe spatial distribution, like statistical techniques that can account for spatial autocorrelation and missing data. Another traditional method is point pattern analysis, which quantifies the degree to which a set of events (i.e. points) are scattered randomly and with differing density across space (Bivand, Pebesma & Gómez-Rubio 2013). A spatial point pattern analysis could be used to model the geographical fragmentation in the spread of one variant for a particular concept, as it would account for both the degree to which the variant is scattered heterogeneously across a particular dialect area and whether particular subregions can be distinguished where the variant is highly frequent. However, to measure the degree of fragmentation for all the variants that occur for a concept, a more aggregate technique is necessary. In disease mapping, less attention has been paid to aggregation techniques for joint disease mapping, which would, in this context, entail modelling the common geographical distribution of more than one disease, as similar patterns are expected for particular environments (e.g. smoking increases the risk of several types of cancer). However, recently, some studies have begun to show interest in these types of research questions (Held et al. 2005). Consequently, comparable aggregate techniques can in the future perhaps also be extended to model the spatial distribution of linguistic variation from an onomasiological perspective. A similar, but map-based, approach has already been employed by Rumpf et al. (2010).

The analysis points to some further open questions. First, the semantic predictor variables, onomasiological vagueness and onomasiological salience, were predominantly measured by relying on the form and distribution of the lexical material available in the database. The degree of affect of a concept was taken into account by relying on aggregated self-reported ratings. However, as the first interaction effect indicates, categorization is not only influenced by the linguistic system that is conventionalized in a speech community, but also by the interaction between this system and the functional needs of the language users (Rosch 1978): some concepts are more locally bound and, for this reason,

29 http://www.nederlab.nl/ (Accessed on 21 August 2017).

seem to be impacted more by the degree of lack of salience. Since we calculate onomasiological vagueness, salience and proneness to affect using supra-local measures (e.g. the total number of multi-word expressions or hapaxes or the total number of non-unique variants in an entire dialect area for a specific concept), we neglect the fact that these semantic features need not be homogeneous throughout the speech community. For particular concepts, the degree of onomasiological salience probably differs between geographical regions or (groups of) speakers, because language users come into contact with various environments in different ways. This is, more specifically, reflected by the fact that the availability of data is geographically limited for some concepts in the database. While for most concepts, this can be explained by the fact that a particular questionnaire of the semantic field of the house was only systematically distributed in the Netherlands (see above), for a second group of concepts, the smaller geographical range is related to the nature of the referents (viz. the Belgian money concepts). Importantly, the degree of onomasiological salience of these concepts also differs geographically: Belgian dialect speakers used the Belgian frank before the introduction of the euro, while people from the Netherlands payed with the Dutch guilder. A second set of concepts that probably have socially varying degrees of onomasiological salience are the concepts related to pigeon keeping. The WBD explicitly mentions that to elicit data for these concepts, they searched for specific informants who were familiar with pigeon keeping, because they have it as a hobby themselves. This indicates that these concepts are probably limited to a particular social group (although the vocabulary collected in part 3 of the dictionaries is expected to be known by every dialect user, Weijnen, Goossens & Goossens 1983: 6). These two examples indicate that the larger dialect regions are also characterized by differences in the micro-socio-cultural and micro-geographical environment of dialect speakers. As a result, it may be the case that social and geographical variation on a smaller scale also impacts the amount of variation in the Brabantic and Limburgish dialects. The relationship between the micro-socio-cultural and micro-geographical environment of a dialect speaker and the language that he uses, will be examined in detail in part 2 of this dissertation.

Second, because this chapter shows that the effect of concept features is stable in other dialect data than were used in the pilot studies, we have obtained further evidence for the impact of semantic concept features on lexical variation in the dialect lexicon at large. However, because dialects are geographically stratified, the extent to which these features influence the geographical distribution of variants or the amount of unique words used for the concept is unclear. Pickl (2013) already indicates that not every aspect of meaning influences both of these aspects in the same way. Furthermore, if semantic concept features not only influence the geographical spread of lexical variants, but also the amount of lexical variation a concept shows, they may be relevant for differently stratified data as well. In the next chapter, we will inquire into the impact of the semantic features distinguished in this chapter on these two dimensions of lexical geographical diversity separately.

4. Deconstructing lexical diversity. An exploratory study

4.1 INTRODUCTION

The previous chapter showed that lexical diversity in dialect data is influenced by characteristics of the concepts under scrutiny. The operationalization of lexical diversity that was used, relies on two aspects of dialectal variation. On the one hand, the number of unique types per concept was taken into account. On the other hand, as dialectal data is stratified along a geographical axis, we also quantified the degree to which the variants for these concepts show a higher degree of spatial fragmentation, by using measures of weighted average dispersion and range per concept. Pickl (2013) indicates that four mechanisms can be distinguished that can explain the extent to which lexical variants are prone to either spatial fragmentation or an increase in the number of unique types. He argues, for instance, that words for concepts that are highly socially relevant, like concepts of salutation with a phatic component of meaning, spread more easily across space, which results in larger and less heterogeneous geographical areas for the lexical variants that occur for these concepts. However, these mechanisms are only discussed in the form of an interpretational framework for the results that he obtains. Consequently, we take a different approach and statistically test the correlation between the concept features and the separate aspects of lexical dialectal diversity.

Inquiring into questions like these is important, because if we can determine *how* concept features interact with different aspects of lexical diversity in dialect data, rather than merely showing that they do, we can, in an exploratory way, assess whether it is possible that they impact lexical variation in differently stratified varieties as well. It may be the case that concept features are also relevant for other lexical data, in the sense that, for instance, between registers, a larger number of different lexical items occurs for specific concepts (cf. DRUNK: *hammered* versus *intoxicated*). Perhaps similar patterns can be found in sociolinguistically stratified data, for example between speakers of different genders or ages, or in diachronic datasets.

In practice, this chapter examines two research questions that concern the extent to which concept features serve as explanatory variables for the different aspects of lexical diversity:

1. Does every concept feature influence the number of unique variants and geographical fragmentation of the concept in the same way and to the same degree?
2. Are the concept characteristics also important if we use an approach that relies on the geographical fragmentation of the lexemes as an explanatory factor of lexical diversity, or did they only reach significance in chapter 3 *because* the data are geographically stratified?

The analysis will consist of two parts, which each aim to provide an answer to one of the questions.

This chapter is structured as follows. Section 4.2 provides a summary of the mechanisms that were distinguished in Pickl (2013) to explain the geographical fragmentation of lexical variants. In 4.3, an overview of the data and methodology used is presented. Section 4.4 outlines the results of the analyses, followed by a discussion and conclusion in 4.5.

4.2 FOUR MECHANISMS OF GEOGRAPHICAL FRAGMENTATION

Preliminary evidence for differences between the effect of concept features on different aspects of lexical diversity comes from Pickl (2013). In Pickl's study, a different methodology is used than the one employed in this dissertation. By means of advanced computational methods, he creates sophisticated dialectological maps that show the spatial

distribution of the variants that occur for a particular concept. On the basis of these maps, the degree to which a concept is characterized by geographical fragmentation is calculated, by relying on measures of both the geographical lack of spread (the inverse of weighted average range, i.e. the extent to which the dialect area is split up into subareas where a particular lexical variant is dominant) and dispersion (the degree to which other variants occur in these subareas of a dominant variant as well) of the lexical variants that are present.[1] Then, for a set of semantic fields, average lack of spread and dispersion per concept are calculated. These results are, finally, interpreted by relying on an explanatory framework that takes into account concept-related features. The four mechanisms that he distinguishes in this framework can be used to devise predictions for the amount of lexical diversity in the Brabantic and Limburgish datasets as well.

According to Pickl, the mechanisms that possibly explain the degree of geographical fragmentation on the dialectological maps, are the following. First, concepts have a high **innovation affinity** if new variants develop regularly. Pickl argues that high innovation affinity occurs often for emotional and expressive concepts (i.e. concepts prone to affect) and causes an increase in the number of unique variants and of the degree of geographical fragmentation (high dispersion and high lack of spread) per concept. For example, for weather phenomena, he indicates that innovation affinity is visible from the difference between the names for wind directions, which show a relatively homogeneous spatial distribution, and more emotionally involved concepts like TO DRIZZLE or TO RAIN HEAVILY, for which more geographical fragmentation is found. The second mechanism is **arbitrariness**. It occurs for concepts with a motivated relationship between *signifiant* and *signifié*, like in the case of onomatopoeic expressions for the sounds of animals. According to Pickl, this mechanism does not directly affect the number of unique types per concept, because arbitrariness can occur for concepts with high or low innovation affinity. It does, however, impact a smaller amount of geographical fragmentation in a speech community, because of an increased chance that the same lexical variant is used in different locations. Third, **diffusion affinity** concerns the extent to which the lexical items for a concept show a high disposition to spread in geographical space, which results in dialect levelling. A high diffusion affinity occurs for concepts that are socially relevant (i.e. across town borders) or that are used by speakers with high mobility, like the salutations with a phatic pragmatic function that were already mentioned above. This mechanism causes a decrease in the geographical fragmentation of a concept (i.e. low degree of dispersion and of lack of spread). The final mechanism distinguished by Pickl is what he refers to as **specificity of meaning**. This mechanism can be interpreted as onomasiological vagueness: "the semantic vagueness can have a massive impact on data elicitation, as the surveyed items may not be represented by a uniform concept in the informants' mental lexicons, resulting in insecure and semantically inconsistent answers, and often multiple responses." (ibid.: 76). For instance, some weather phenomena, like DRIZZLE, HEAVY RAIN, SLEET, HAIL and SNOWSTORM, form an onomasiological continuum and are difficult to distinguish from each other. Following Pickl's argumentation, we can, therefore, expect that onomasiological vagueness predominantly affects the number of unique types per concept, because the chance that different language users make the same demarcation choices is smaller. Furthermore, he asserts that it only affects the average dispersion for a concept, but that it is less important for the average geographical range of the variants used.

On the basis of the mechanisms that Pickl distinguishes, three hypotheses can be formulated[2]:

1. **diffusion affinity**: concepts that are less relevant in supralocal communication, as visible from the fact that they belong to **semantic fields** that are locally bound, will, on average, show a higher degree of geographical fragmentation (i.e. higher weighted average dispersion and higher weighted average lack of spread) per concept. Additionally, although Pickl does not mention it explicitly, we can expect that this mechanism also causes an increase in the number of unique types per concept, because a high affinity to diffusion in space causes a higher degree of dialect levelling;
2. **specificity of meaning**: the degree of **onomasiological vagueness** of a concept will predominantly correlate positively with a larger number of available unique lexical items and with a higher average geographical dispersion per concept. Weighted average lack of spread per concept will be less affected by this variable;
3. **innovation affinity**: the degree of **sensitivity to affect** of a concept will simultaneously correlate positively with the number of unique lexical items per concept, the weighted average dispersion per concept, and the weighted average lack of spread per concept.

1 The calculation of lack of spread and dispersion in Pickl (2013) differs from the operationalizations used in this dissertation, but the underlying rationale is highly comparable. Pickl also uses a different terminology than 'range'/'lack of spread' (viz. 'complexity') and 'dispersion' (viz. 'homogeneity'), but to avoid confusion, we do not adopt his terminology here.

2 Pickl's notion of arbitrariness is not included in our analyses.

	number of unique types	weighted average dispersion	weighted average lack of spread
locally-bound semantic fields (vs. supra-local)	+?	+	+
onomasiologically more vague concepts (vs. non-vague)	+	+	/
affect-sensitive concepts (vs. neutral)	+	+	+

TABLE 4.1
Summary of the hypotheses distinguished on the basis of Pickl (2013); no hypotheses are available for onomasiological salience

These hypotheses are summarized in Table 4.1. A plus sign '+' indicates that a positive correlation is expected and a forward slash '/' means that the type of lexical diversity will probably not be influenced by the concept feature. The first column outlines the concept features for which predictions are available (no hypotheses have been discerned for onomasiological salience). The second column summarizes the hypotheses for 'number of unique types'. As Pickl does not explicitly mention any predictions for the relationship between locally-bound semantic fields and the number of unique types, the table shows '+?'. The second column shows the expectations for 'weighted average dispersion'. The final column gives the predictions for 'weighted average lack of spread', the degree to which the variants for the concept are limited to a small geographical region.

4.3 DATA & METHODOLOGY

In this chapter, we use two ways of analysing variability in the response variables to determine whether we can confirm (1) that the effect of the concept features can differ between concept features and (2) that they remain relevant if we use geographical fragmentation as an explanatory factor of lexical diversity. The analyses are conducted on the dataset used in chapter 3 to ensure that the results are comparable. Thus, overall, the datasets consists of 3136 concepts from the WLD and WBD, collected from six volumes (i.e. semantic fields) from the dictionaries. The following sections provide an overview of the dependent (4.3.1) and independent variables (4.3.2) used in the analyses and of the methodology that we employ (4.3.3).

4.3.1 Dependent variables

In the first part of the analysis, we examine research question 1, whether the concept features influence the number of unique variants and geographical fragmentation of the concept in the same way and to the same degree. More specifically, we use the three separate aspects of the response variable that were already discussed in chapter 3. The first variable concerns **the number of unique types** available per concept. The second and third variable each model an aspect of a concept's degree of **geographical fragmentation**. More specifically, the second variable, **(weighted average) dispersion**, gauges the weighted average of the geographical distance between two locations with the same variant for a particular concept. The third variable, **(weighted average) range**, concerns the weighted average of the proportion of the geographical surface where a particular lexical variant occurs per concept. In the analyses, we use the inverse of this variable, which we refer to as **(weighted average) lack of spread**, because if, on average, the lexical items used for a particular concept occur in a smaller geographical area (i.e. they have a higher degree of lack of spread), the concept shows more geographical fragmentation. The calculation of these variables was presented in chapter 3. Table 4.2 provides an overview of the distribution of these variables, with one outlier for lack of spread excluded (viz. VERSCHILLENDE KNIKKERSPELEN 'various games of marbles'; lack of spread = 347.70). In practice, we

variable	distribution
unique number of types	minimum = 1, maximum = 202, mean = 23.30, sd = 20.59
(weighted average) dispersion	minimum = 1, maximum = 4.40, mean = 1.89, sd = 0.59
(weighted average) lack of spread	minimum = 1, maximum = 60.82, mean = 2.30, sd = 2.46

TABLE 4.2
Overview of the dependent variables in the first part of the analyses

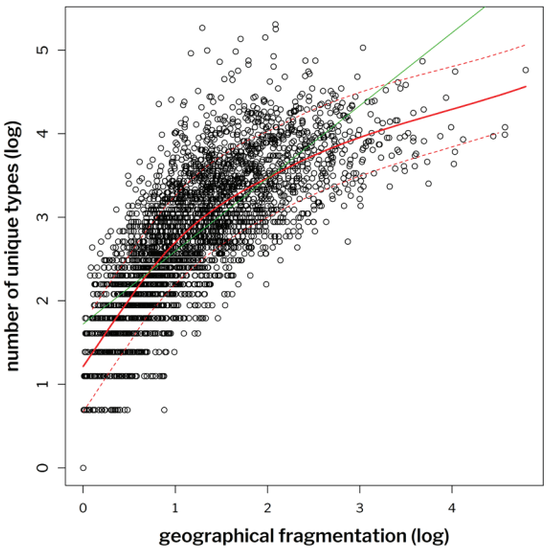


FIGURE 4.1
The relationship between geographical fragmentation and number of unique types

use the natural logarithm of these variables in the analyses to increase the degree of linearity between the explanatory and dependent variables.

In the second part of the analyses, we conduct an exploratory study of the extent to which concept features remain influential for the number of unique types that are available per concept if we account for the geographical signal in the dialect data (research question 2). Before explaining the method that we use to take this signal into account, we briefly examine the relationship between the number of unique types per concept and the amount of geographical fragmentation a concept shows, i.e. dispersion divided by range (which is equal to the product of dispersion and lack of spread). Figure 4.1 visualizes the degree to which these variables are correlated.

The figure indicates that, on average, a higher degree of geographical fragmentation (on the x-axis) correlates with a higher number of unique types per concept (on the y-axis; Spearman's rho = 0.79, $p < 0.001$). However, as the red curved line indicates, the relationship is only linear-like: for some concepts that show a high amount of geographical fragmentation, fewer unique types are available than predicted. This is, for instance the case for MINACHTEN 'to despise' in the WBD, a concept that occurs with a moderately large amount of lexical items in the dataset (viz. 44), but that shows a high degree of geographical fragmentation, with dispersion equal to 2.963 and lack of spread equal to 19.01. The difference between the number of unique types and geographical fragmentation for this concept can be explained by the fact that none of the lexical items are frequent across the Brabantic dialect area: 36 of the variants occur only once or twice and the most frequent variant is only available 8 times. Furthermore, for many concepts with an average degree of geographical fragmentation, more unique types are available than expected. These concepts are located above the green line. BUIK (SPOTNAMEN) 'jocular names for the stomach' in the WLD, for example, has an average value for dispersion (2.26) and a low value for lack of spread (1.50), but occurs with an above-average number of unique types in the data (61). For this concept, one lexical item, *pens*, occurs 246 times in the dataset and it is used throughout the Limburgish dialect area, which indicates that the concept shows a relatively large degree of homogeneity: throughout the dialect area, the language users employ the same word to refer to the concept. However, occasionally and in more spatially limited regions, they also use alternative lexical items. These latter lexemes are much less frequent - the second most frequent term for this concept, *buikje*, occurs only 18 times - and as a result, they contribute less to the weighted calculation of the average geographical dispersion and lack of spread for the concept.

The second part of the analyses then examines whether differences like the ones we find between concepts like MINACHTEN and BUIK (SPOTNAMEN) are also affected by the concept features that were distinguished in chapter 3, because this serves as an exploratory way of determining whether the concept features only reach significance *because* the data are geographically stratified. The operationalization of the response variable is as follows. First, we build a linear regression model with as a predictor the natural logarithm of geographical fragmentation per concept (i.e. dispersion divided by range) and as the response the natural logarithm of the number of unique types per concept. The residuals of this model represent the variance in the number of unique types that remains unexplained if the geographical signal is accounted for. The model used to obtain these residuals has an adjusted R^2 value of 0.5621, which indicates that about 56% of the variance in the number of unique types can be explained by the geographical fragmentation of the concepts. The model diagnostics do not reveal problems with outliers or heteroscedacity. The relationship between the predictor and the response is linear-like [3] and the residuals are very close to normally distributed.

3 We also examined whether using the residuals of a linear regression analysis that models the impact of the logarithm of geographical fragmentation on the number of unique lexical items per concept, without taking the logarithm of the latter variable, provides a better fit of the model to the data. Although the relationship between geographical fragmentation and number of unique types is more linear with this procedure, this model performs worse, because it suffers from a large amount of outliers and from heteroscedacity due to the fact that the variable 'number of unique types' is highly skewed: only a few concepts occur with very high values for this variable.

independent variable	possible values / distribution	
dictionary	WLD WBD	N = 1456 N = 1680
semantic field	the human body the house celebration & entertainment personality & feelings family & sexuality society, school & education	N = 361 N = 508 N = 471 N = 703 N = 119 (WLD only) N = 974
lack of salience		
proportion of MWE's	minimum = 0, maximum = 1, mean = 0.12, sd = 0.22	
proportion of hapaxes	minimum = 0, maximum = 0.79, mean = 0.09, sd = 0.09	
prevalence binary	prevalent missing / not prevalent	N = 1791 N = 1345
vagueness		
lexical non-uniqueness	minimum = 0, maximum = 257, mean = 16.76, sd = 25.15	
affect		
proportion of non-neutral ratings	minimum = 0, maximum = 1, mean = 0.43, sd = 0.43	

TABLE 4.3
Overview of the independent variables

Then, we use these residuals as the response variable in a linear regression model for the second part of the analysis. This residualized response variable takes values between -1.80 and 2.42 with mean 0 and standard deviation 0.57. For MINACHTEN, for instance, the residuals are negative (-1.45), while for BUIK (SPOTNAMEN), the residuals are positive (1.32). This response variable thus represents whether a particular concept occurs with more or fewer unique lexical variants, given the amount of geographical fragmentation of these variants. For this reason, geographical fragmentation also serves as an explanatory variable. Importantly, however, by using this methodology to operationalize the response variable, we do not take into account the degree of homogeneity in the (geographically stratified) profile of the concepts, i.e. the degree to which a particular variant (or a few variants) takes precedence over alternative expressions in the dialect area. Instead, we merely inquire into the number of equivalent, synonymous expressions that exist in a particular speech community, regardless of how frequent these synonyms are. Other ways to account for the geographical signal may therefore yield different results. However, as, to the best of our knowledge, we are the first to examine the influence of concept features on lexical diversity on a large scale and in a systematic way, we believe that, as a first case-study, our methodology will uncover possibilities for improvement and new lines of research.

4.3.2 Independent variables

To ensure maximal comparability, we use the same independent variables that were included in the analyses in chapter 3. However, we exclude the proportion of missing places per concept, as this variable was shown to be problematic. Additionally, from a theoretical perspective, the explanatory variable 'proportion of hapaxes' is strongly correlated with the response variable 'number of unique types per concept', because for two concepts with an identical number of tokens, a larger amount of hapaxes directly correlates with the number of unique types available for the concept.[4] For

4 In practice, the number of tokens per concept takes values between 34 and 1496, with mean 169.7 and standard deviation 89.36, which indicates that including the proportion of hapaxes is not as problematic as we assume here. Furthermore, it could be argued that a higher number of available tokens also elucidates concept-related features, like the fact that the concept is more onomasiologically salient. Although the proportion of hapaxes shows a significant strong positive correlation with the number of unique types per concept, this relationship is not linear (spearman's rho = 0.833, $p < 0.001$).

this reason we do not include the proportion of hapaxes as a predictor in the model for the number of unique types per concept and to ease comparability, it is also excluded from the other models in the first part of the analysis. However, it is less strongly linked to the weighted average dispersion and lack of spread, the other response variables used in the first part of the analysis, because to calculate these variables, a weighting scheme is used that takes into account the profile of the concept (see chapter 3). Table 4.3 provides an overview of the independent variables.

4.3.3 Methods

The first part of the analysis examines whether the concept features have the same effect on the three dependent variables and whether the effect sizes are the same. We build three separate multiple linear regression models. The effect of the predictors on the number of unique types per concept is measured in model 1, on the weighted average dispersion per concept in model 2 and on the weighted average lack of spread per concept in model 3. After verifying that the direction of the effect of the predictors is the same in all three models and after checking the assumptions of the models, we inquire into the relative contribution of each predictor to the separate models using type II analyses of variance. Because we are mostly interested in comparing the relative contribution of the main effects of the predictors between the three models, we do not include interaction effects. However, for each model, we verified that the effect of the concept features does not differ between the dictionaries or semantic fields. We use the same formula for all three models:

log(*response variable*) ~ dictionary +
semantic field +
lack of salience -- proportion of MWE's +
lack of salience -- prevalence binary +
vagueness -- lexical non-uniqueness +
affect -- proportion of non-neutral responses

In the second part of the analysis, we conduct an exploratory analysis of whether the impact of the concept features holds if we account for the geographical signal in the data. As explained above, we use the residuals of a linear regression model that gauges the effect of geographical fragmentation on the number of unique types per concept, as the response variable in our final linear regression model. For variable selection, we use an automatic backward selection procedure, based on a comparison of the AIC criterion of the models, while allowing for two-way interactions between all the predictors. We subsequently exclude model terms and interactions that do not significantly contribute to the explanatory power of the model. Additionally, we remove interactions between single predictors that gauge the same concept feature to avoid a harmful amount of multicollinearity. Finally, we avoid overfitting by only retaining the interactions that contribute enough to the reduction of the variance in the response variable. Before interpreting the results of the model, the assumptions of the model are examined. More specifically, we verify that the relationship between the predictors and the response is linear, that there are no outliers, that the model does not suffer from heteroscedacity, that there is no harmful multicollinearity between the predictors and that the residuals are normally distributed.

4.4 RESULTS

4.4.1 Comparing the relative contribution of the concept features

This section provides an analysis for the first research question: does every concept feature influence the number of unique variants and geographical fragmentation of the concept in the same way and to the same degree or can we establish that, as the framework outlined in Pickl (2013) indicates, different concept features may have different effects? More specifically, we examine the relative impact of the concept features on the unique number of variants (model 1), weighted average dispersion (model 2) and weighted average lack of spread per concept (model 3).[5]

First, the output of the models, shown in Table 4.4, indicates that all the predictors have the expected effect: we consistently find a positive correlation between all three response variables (i.e. more lexical diversity) and the lack of onomasiological salience, degree of onomasiological vagueness and proneness to affect. Additionally, in the WBD, the number of unique types and the weighted average dispersion per concepts is higher, but the weighted average lack of spread is significantly lower. This is probably related to geographical differences between the Brabantic and Limburgish dialect areas. Because the Brabantic region is larger, more densely populated and because the Brabantic dataset contains more records, it is easier to find a larger number of different lexical items that are more dispersed. Additionally, it results in a significantly lower average value for lack of spread (i.e. the inverse of range) per concept.

5 The significance of the predictors on the response variables were also verified using multivariate regression analysis, which allows for modelling the impact of predictors on several response variables at the same time (Field, Miles & Field 2012: 696-738, Fox & Weisberg 2011). Using Pillai's trace we found that all the predictors reach significance in the multivariate model.

model term	model 1 number of unique types			model 2 weighted average dispersion			model 3 weighted average lack of spread		
	estimate	SE	p-value	estimate	SE	p-value	estimate	SE	p-value
intercept	2.293	0.038	< 0.001	0.402	0.015	< 0.001	0.186	0.026	< 0.001
dictionary									
WBD	0.141	0.024	< 0.001	0.055	0.009	< 0.001	-0.155	0.016	< 0.001
semantic field									
the house	0.052	0.044	NS	0.035	0.017	< 0.05	0.169	0.030	< 0.001
celebration & entertainment	-0.191	0.044	< 0.001	0.010	0.018	NS	0.145	0.031	< 0.001
personality & feelings	-0.078	0.045	< 0.1	0.015	0.018	NS	0.044	0.031	NS
family & sexuality	-0.122	0.068	< 0.1	0.043	0.027	NS	-0.001	0.047	NS
society, school & education	-0.143	0.039	< 0.001	0.040	0.015	< 0.05	0.036	0.027	NS
lack of salience									
proportion of MWE's	0.933	0.056	< 0.001	0.350	0.022	< 0.001	0.689	0.038	< 0.001
prevalence binary (missing / not prevalent)	0.233	0.025	< 0.001	0.092	0.010	< 0.001	0.196	0.018	< 0.001
vagueness									
lexical non-uniqueness	0.015	0.001	< 0.001	0.004	0.000	< 0.001	0.006	0.000	< 0.001
affect									
proportion of non-neutral ratings	0.463	0.033	< 0.001	0.088	0.013	< 0.001	0.144	0.022	< 0.001

TABLE 4.4
Output of the three linear regression models

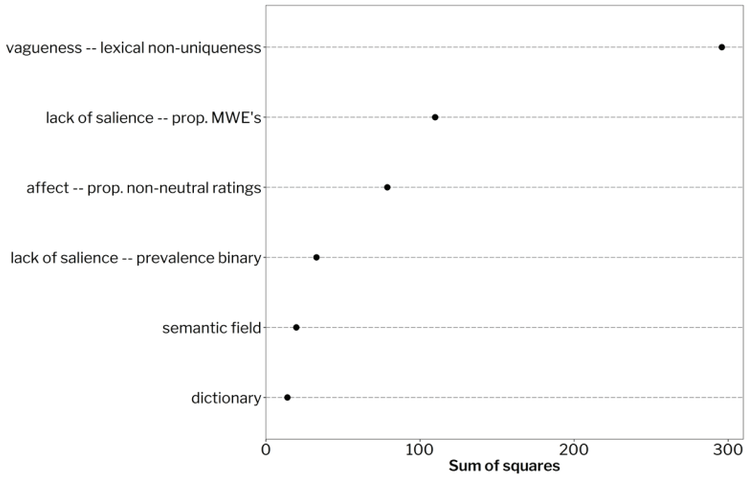


FIGURE 4.2
Analysis of variance for model 1 – number of unique types per concept

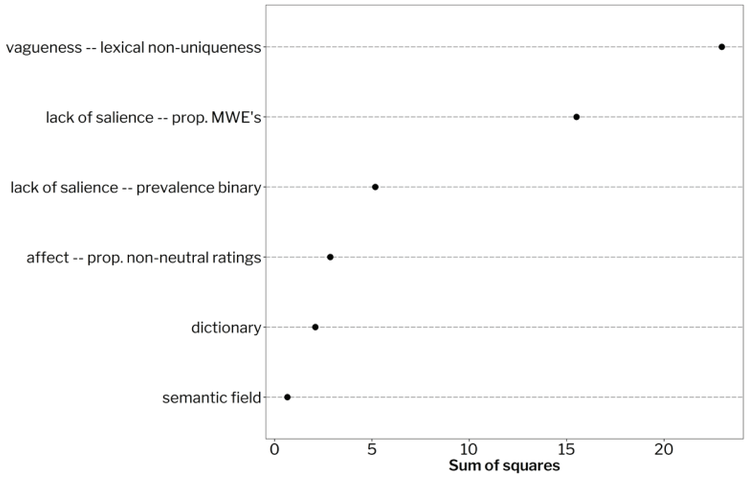


FIGURE 4.3
Analysis of variance for model 2 – weighted average dispersion per concept

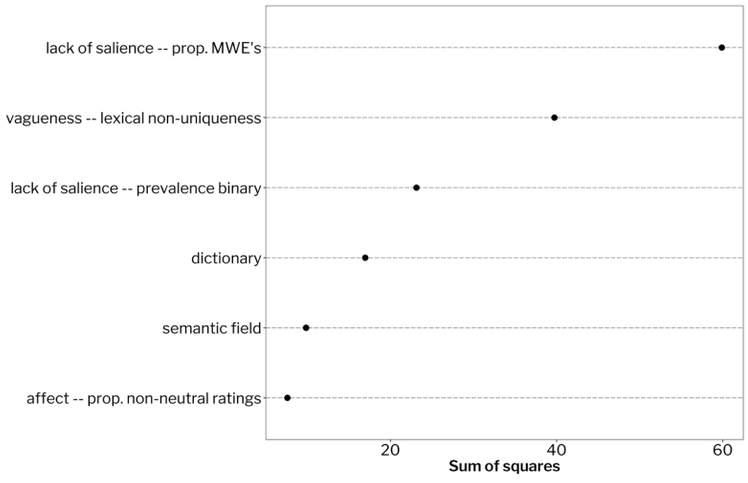


FIGURE 4.4
Analysis of variance for model 3 –weighted average lack of spread per concept

We also find diverging patterns between the semantic fields. More specifically, concepts belonging to the society-related semantic fields 'celebration and entertainment' and 'society, school & education' occur with significantly fewer unique types in the data, in comparison to the reference level, 'the human body'. However, on average, these lexical items also have a significantly higher dispersion or lack of spread. Additionally, concepts of the semantic field 'the house' show significantly more dispersion and a higher lack of spread than human body concepts.[6] These findings are only partly in line with the hypothesis outlined in 4.2, that less locally bound semantic fields show more lexical levelling. However, semantic field does not contribute much to the explanatory power of any of the models and it is only borderline significant in model 2 (average dispersion per concept). Adjusted R^2 is 0.48 in model 1, it reaches 0.33 in model 2 and 0.30 in the third model.

Figures 4.2 to 4.4 show the relative impact of the concept features in the three models. More specifically, the sums of squares explained by each of the variables in the models, obtained with type II analyses of variance, are plotted. These sums of squares represent the extent to which each predictor contributes to explaining the variance in the dataset. The x-axis shows the absolute value of the sums of squares. However, as these values are model-dependent, we will only interpret them in a comparative way by taking into account the relative order of the predictor variables (on the y-axis).

The figures indicate that, overall, the influence of the explanatory variables is comparable between the models. However, some minor differences occur. For the number of unique types (model 1), lexical non-uniqueness, a measure of onomasiological vagueness, is clearly the most important predictor. Two other predictors, proportion of multi-word expressions and proneness to affect, also explain a relatively large amount of the variance, while the impact of the other predictors is smaller. In model 2, lexical non-uniqueness is the most important variable as well, but the impact of the proportion of multi-word responses is not much smaller. Furthermore, affect is less important than in model 1. In model 3, finally, the proportion of multi-word responses is the most influential predictor, which may indicate that, in comparison to the other explanatory variables, onomasiological salience is more important to explain the lack of spread of a concept. This is also confirmed by the fact that the other measure of lack of salience, prevalence binary, also explains a relatively large amount of variance in comparison to its effect in the other models. Lexical non-uniqueness takes up the second position in model 3. Interestingly, proneness to affect does not play a large role in this model. The figures also show that the least important variables are nearly always the same in the three models: differences between the dictionaries and semantic fields are not of great importance for any of the response variables. This indicates that the concept features affect each aspect of lexical geographical heterogeneity to some extent.

The difference between the three figures is predominantly related to measures of onomasiological salience, affect, and onomasiological vagueness. First, onomasiological salience seems to be more important for the weighted average dispersion and lack of spread of a concept, but the effect of this variable is smaller for the number of unique lexical variants. Dispersion and lack of spread take a low value for concepts for which one variant is very frequent throughout a dialect area (in comparison to other variants that may occur) and for which it also holds that this frequent variant shows a relatively uniform spatial distribution. The variables take a high value for concepts for which many lexical items occur frequently, that, on average, also occur in a relatively irregular way in a limited region. Perhaps the effect of onomasiological salience is, therefore, more important for these response variables because a higher degree of salience causes a more homogeneous geographical profile. One could, for instance, imagine that for more salient concepts, more clear geographical areas where one variant is highly frequent, are found. An extreme example, for instance, is BLOED 'blood', a highly salient concept for which one variant, *bloed*, is used throughout the Brabantic and Limburgish dialect areas.

On the basis of the relative ranking of the predictors in Figures 4.2-4.4, we can furthermore determine whether we find evidence for the hypotheses outlined in 4.2, which are based on the explanatory framework of Pickl (2013). Recall that, as Table 4.1 showed, we expected to find that differences between semantic fields affect all three aspects of lexical diversity (although the effect on the number of unique types was not explicitly mentioned by Pickl), that onomasiological vagueness influences the number of unique types and the weighted average dispersion and that affect-sensitivity correlates with all three aspects of lexical diversity.

The comparison of the models confirms the prediction for onomasiological vagueness, (i.e. the fourth mechanism distinguished by Pickl, specificity of meaning). While onomasiological vagueness reaches significance in all three models, comparing the relative order of the variables in the figures indicates that it has the largest impact on the

6 As we are examining the same dataset as was used in chapter 3, the chance of encountering type I errors increases. As a result, effects that do not have a very small p-value may not all be reliable. However, as the influence of semantic field does not play a large role in the analysis, this will not cause any problems for the interpretation of the models. Furthermore, all the type II analyses of variance indicate that the predictors included in the models have very low p-values that approach 0.

number of unique lexical items per concepts, as it is clearly set off from the other variables in Figure 4.2 (model 1). It also takes up the first position in Figure 4.3 (model 2), but the impact of the second-ranking predictor, proportion of multi-word expressions, is not much smaller. Consequently, we can infer that while vagueness causes a larger number of unique lexemes to be used, which are dispersed across space in a heterogeneous way, some of these lexical items occur in a relatively large region, as exemplified by the fact that the correlation between vagueness and lack of spread is less strong. These findings can be explained by taking into account that demarcational differences may be made by different speakers. First, for vague concepts, many lexical items are available due to such demarcational differences. For instance, for the vague concept MOKKEN, in Rosmalen (in the north of the province of North Brabant), the same lexical item *pratten* is used for TREUREN as well. In Berghem, however, a close-by location also in the north of the province of North Brabant, a different lexical item is available which also has other meanings: *grijnzen* occurs for MOKKEN, but it can also be used for HUILEN 'to cry' and DRENZEN 'to whine'. However, the geographical distribution of some of these lexical items may be relatively large because speakers from geographically distinct locations can still make the same choices regarding the demarcation of the onomasiologically vague concepts. The spread of these lexical items is, however, highly irregular, as indicated by the larger effect of vagueness on dispersion.

However, some of Pickl's predictions are contradicted by the figures as well. First, we hypothesized that affect would influence all aspects of lexical geographical heterogeneity due to the large innovation affinity of expressive and emotionally involved concepts. Nonetheless, the figures indicate that it is mostly relevant for the number of unique types and, to a lesser extent, for the weighted average dispersion per concept. The impact of affect-sensitivity on the average geographical lack of spread of a concept is very small. An explanation for this finding may be that, while the affect-sensitive concepts are prone to lexical innovation, perhaps for some concepts, this does not completely prevent the spatial distribution of these innovative lexemes. On the one hand, a less strong correlation between affect and lack of spread can occur if for some affected concepts, new variants still diffuse relatively quickly across geographical space, because innovative ways to refer to the concepts are borrowed throughout the dialect areas. The stronger correlation between affect and dispersion can then be explained by the fact that in many locations, more than one synonym to refer to the affected concept is probably available. On the other hand, perhaps people from different locations sometimes use the same source domain to coin new words for the affect-sensitive concepts, which can result in a less clear correlation between affect and lack of spread.[7]

Furthermore, the prediction regarding the impact of semantic field, which coincides with Pickl's third mechanism, diffusion affinity, is also not confirmed. We hypothesized that semantic field influences all aspects of lexical diversity, because the speed with which a particular variant spreads across geographical space, would differ between the locally-bound and supralocal semantic fields. However, the figures indicate that differences between semantic fields only play a minor role in all three models. Furthermore, as outlined above, the effects that do reach significance only partly corroborate Pickl's assumptions.

4.4.2 Exploring the influence of the concept features while accounting for the geographical signal

In this section, we present a way of analysing whether the effect of the concept characteristics is preserved if we use an approach that relies on the geographical fragmentation of the lexemes as an explanatory factor of lexical diversity. More specifically, it allows us to examine, in an exploratory way, whether the concept features only reach significance *because* the data are geographically stratified (research question 2).

As outlined above, the response variable in the linear regression model discussed in this section consists of the residuals of a linear regression model that gauges the impact of geographical fragmentation (i.e. weighted average dispersion divided by weighted average range) on the number of unique lexical items per concept. The Adjusted R^2 value of the final model is 0.37, which indicates that 37% of the variance in the response variable can be explained by the concept features. Table 4.5 shows the impact of the significant concept features on the residualized response variable.

7 Although the results for dispersion and affect are to some extent comparable, the argument outlined here suggests that these results are influenced by different phenomena. On the one hand, the larger weighted average dispersion of some vague concepts is argued to result from the fact that different demarcational choices need not depend on the location of the speaker and are not conventionalized across the speech community: although a particular demarcational choice may be made in one location, this does not entail that the same choice is made in a nearby location. As a result, we implicitly assume that vaguer concepts are not very frequent in discourse across town borders. On the other hand, the larger weighted average dispersion for affected concepts, is asserted to result from the fact that affect-sensitive concepts are under communicative pressure, probably because they are more salient in discourse. This then results in a large affinity to innovation and creativity, which causes a larger degree of dispersion. However, although the interpretations seem to be in accordance with the mechanisms outlined by Pickl, the frequency differences between onomasiologically vague and affected concepts in supralocal discourse need to be confirmed by means of other types of linguistic data (e.g. by using appropriate corpora).

model term	estimate	SE	p-value
intercept	-0.4420	0.0360	< 0.001
dictionary			
WBD	0.2120	0.0170	< 0.001
semantic field			
the house	0.0130	0.0460	NS
celebration & entertainment	-0.0750	0.0460	NS
personality & feelings	0.2060	0.0510	< 0.001
family & sexuality	-0.0940	0.0700	NS
society, school & education	-0.0730	0.0410	< 0.1
lack of salience			
proportion of hapaxes	2.3460	0.3630	< 0.001
vagueness			
lexical non-uniqueness	0.0220	0.0020	< 0.001
affect			
proportion of non-neutral ratings	0.2070	0.0240	< 0.001
interaction terms			
sem. field (*the house*) : proportion of hapaxes	-1.2440	0.4940	< 0.05
sem. field (*celebration & entertainment*) : proportion of hapaxes	-2.3410	0.4190	< 0.001
sem. field (*personality & feelings*): proportion of hapaxes	-1.8670	0.4040	< 0.001
sem. field (*family & sexuality*) : proportion of hapaxes	1.2250	0.7220	< 0.1
sem. field (*society, school & education*) : proportion of hapaxes	-0.6110	0.4130	NS
sem. field (*the house*) : lexical non-uniqueness	-0.0020	0.0020	NS
sem. field (*celebration & entertainment*) : lexical non-uniqueness	0.0000	0.0030	NS
sem. field (*personality & feelings*): lexical non-uniqueness	-0.0080	0.0020	< 0.001
sem. field (*family & sexuality*) : lexical non-uniqueness	-0.0140	0.0060	< 0.05
sem. field (*society, school & education*) : lexical non-uniqueness	-0.0060	0.0020	< 0.05
proportion of hapaxes : lexical non-uniqueness	-0.0570	0.0040	< 0.001

TABLE 4.5
Output of the linear regression model for the residualized response variable

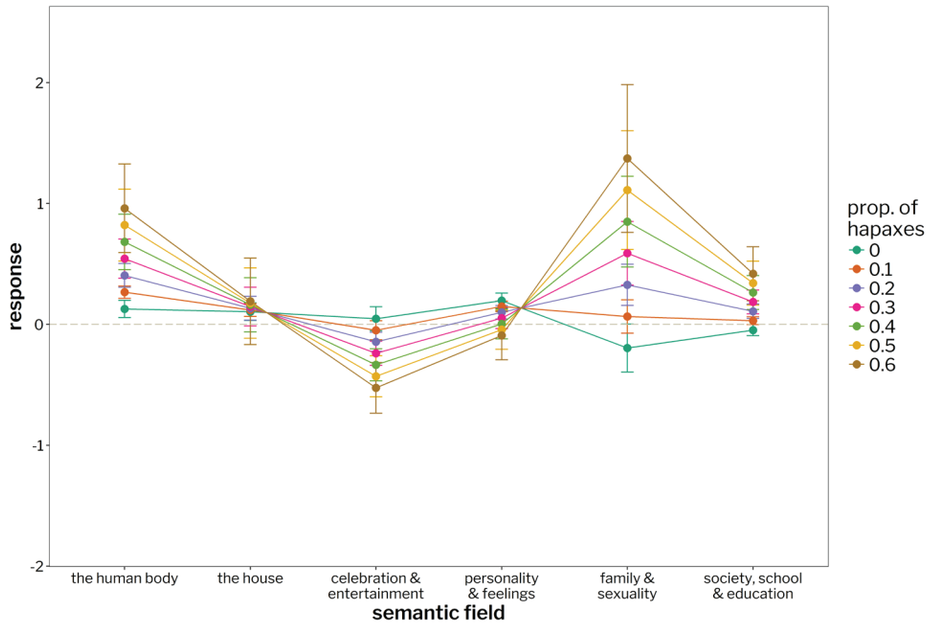


FIGURE 4.5
Interaction between semantic field and proportion of hapaxes

Positive values for the estimates indicate that more unique types per concept than predicted are available, given the amount of geographical fragmentation the concept shows (i.e. the residual variance is larger than predicted). The estimates take negative values if fewer unique types occur than predicted, given the amount of geographical fragmentation for the concept (i.e. the residual variance is smaller than predicted).

For the first main effects predictor that is not included in any interactions, dictionary, the table indicates that in the WBD, the number of unique types is generally larger than in the WLD, given the amount of geographical fragmentation and all other things being equal. This finding is probably related to geographical differences between the dictionaries: the concept-based measurements for the WBD are based on a larger number of observations (see chapter 3: 328 320 versus 204 307 in the WLD). Crucially, there are no significant differences concerning the effect of any of the other predictors between two dictionaries.

For the second predictor that is not included in any interaction effect, proportion of non-neutral ratings per concept, the table shows that concepts that are more prone to affect occur with significantly more unique types, given the amount of geographical fragmentation that the concepts show and all other things being equal. Consequently, affect-sensitivity extends beyond the geographical profile of the concepts: the amount of lexical diversity for affect-sensitive concepts in these dialectal data cannot solely be explained by the fact that the concepts show a certain degree of geographical fragmentation.

The interaction effects are best interpreted visually. Note that all the interaction effects were also included in the model discussed in chapter 3, which already suggests that these results are stable even when the geographical signal is accounted for. Figure 4.5 shows the effect of the interaction between the semantic field and the proportion of hapaxes.[8] A grey dashed line shows where the response is equal to 0: above this line, more unique types occur than expected, given the amount of geographical fragmentation the concept shows, while below this line, fewer unique types are available than expected on the basis of the amount of fragmentation. The model mostly indicates that the impact of proportion of hapaxes, a measure of onomasiological salience, is significantly smaller in specific semantic fields, viz. 'the house', and 'personality & feelings', in comparison to the reference level, 'the human body'. While the proportion of hapaxes seems to have a larger effect on the semantic field of family & sexuality, this effect is only borderline significant ($p < 0.1$) in comparison to the reference level.

8 As outlined above, it may be the case that the proportion of hapaxes directly correlates with the number of unique types per concept. Furthermore, one could imagine that another explanation for the results shown in Figure 4.5 is that the effect of proportion of hapaxes is larger in particular semantic fields merely because some fields show a higher degree of geographical fragmentation in general, which may reduce the potential effect of the number of unique types per concept. However, this explanation does not hold. Only small differences occur (Adjusted $R^2 = 0.06$, $p < 0.001$) and they do not follow the pattern predicted by the model. Consequently, we can conclude that the importance of the proportion of hapaxes differs between semantic fields.

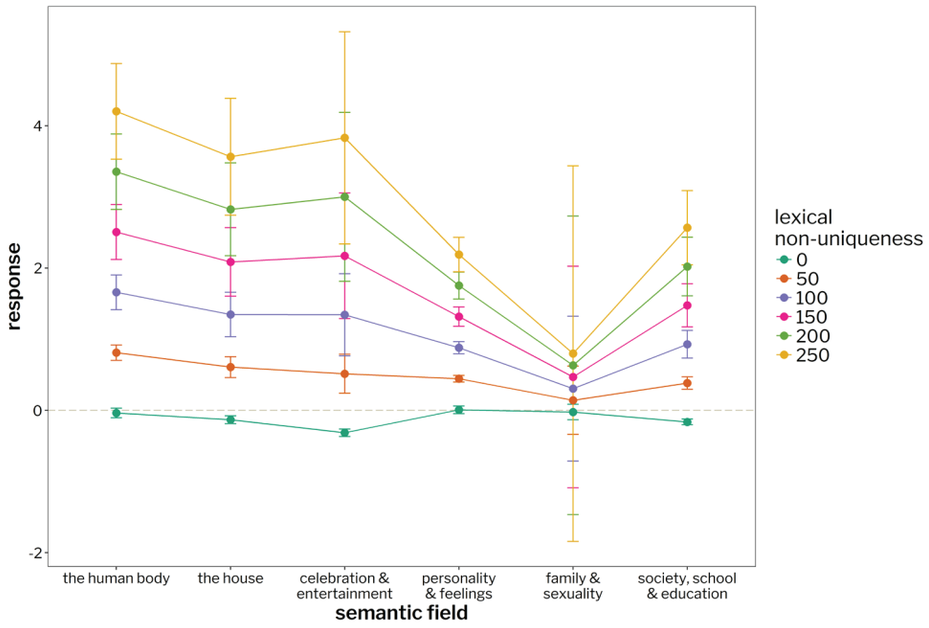


FIGURE 4.6
Interaction between semantic field and lexical non-uniqueness

Additionally, if the predictor has an effect, it is generally the case that less salient concepts show a larger number of unique types, given the amount of geographical fragmentation. Although the figure indicates that the opposite effect is found for the semantic fields 'celebration & entertainment' and 'personality & feelings', the difference between concepts with diverging proportions of hapaxes is very small, as indicated by the error bars. Consequently, we can assume that the degree of onomasiological salience impacts the number of lexical types per concept to some extent, when the geographical signal is accounted for. However, Table 4.5 shows that none of the other variables that gauge the degree of salience of a concept reach significance in a multifactorial environment. As a result, the impact of onomasiological salience beyond the geographical profiles of the concepts is only limited.

The second interaction effect concerns the relationship between semantic field and lexical non-uniqueness, which gauges the extent to which a concept is vague. Recall that in chapter 3, it was argued that the interpretation for concepts with a very high proportion of non-unique types is not as reliable as such concepts are not very frequent in the dataset. Additionally, the results for the semantic fields 'celebration & entertainment' and 'family & sexuality' are problematic as these semantic fields only contain few very vague concepts.

Figure 4.6 shows the visual representation of the interaction effect. It indicates that, as expected, for vaguer concepts, more unique types per concept are available than predicted given the geographical fragmentation of the concepts and all other things being equal. Furthermore, the effect of lexical non-uniqueness seems to differ between the concrete and abstract semantic fields. For the semantic fields with more abstract concepts, personality & feelings, family & sexuality and society, school & education, towards the right of the plot, the impact of lexical-non uniqueness is significantly smaller than for the reference level, the human body. The difference between the reference level and the other concrete semantic fields, the house and celebration & entertainment, on the left side of the plot, is not significant. The interpretation for these finding is perhaps again related to the fact that for abstract concepts, fewer perceptual clues are available, which may result in a reduction of the potential effect of lexical non-uniqueness for these concepts. Overall, these results seem to indicate that, even when the geographical signal is accounted for, vaguer concepts occur with a larger number of unique types than expected, although this effect is more outspoken for concrete concepts. Consequently, the effect of lexical non-uniqueness on the amount of lexical diversity in the dialect data cannot solely be attributed to the geographical fragmentation of the concepts.

The third interaction effect models the combined effect of lexical non-uniqueness and proportion of hapaxes on the response variable. Figure 4.7 indicates that for concepts with a small proportion of hapaxes, lexical non-uniqueness correlates positively with the number of unique types that are available, given the geographical fragmentation of the concept. More specifically, for salient concepts that are vague, a larger number of different types occurs per concept than for salient concepts that are not vague, given the degree of geographical fragmentation. For concepts with a larger

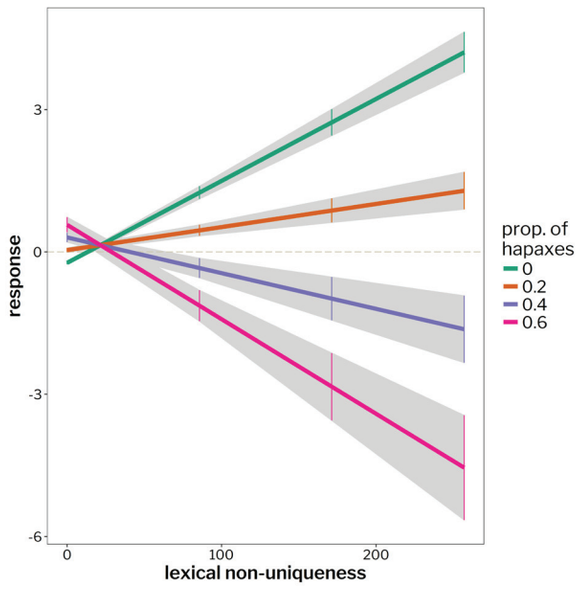


FIGURE 4.7
Interaction between lexical non-uniqueness and proportion of hapaxes

proportion of hapaxes, i.e. less salient concepts, the impact of lexical non-uniqueness decreases and the correlation is even negative for concepts with a very high proportion of hapaxes. However, this finding needs to be attenuated. More specifically, most of the concepts have a relatively low proportion of hapaxes: only 184 out of the 3136 concepts in the database have a proportion of hapaxes larger than 0.25. Consequently, for the highly non-salient concepts, the effect plot is only based on a small number of observations.

In chapter 3, we asserted that differences in vagueness may predominantly affect highly salient concepts (e.g. SLIM (ZIJN) 'to be smart' versus BLOED 'blood' in the WBD). Non-salient concepts, however, show a lot of variation in general, which reduces the possible effect of onomasiological vagueness (e.g. STAARTDUIF 'pigeon listed as one of the last winners (pigeon keeping)' a non-salient, but also non-vague concept that still shows a lot of lexical diversity in the WBD). The high degree of diversity for non-salient concepts may be related to the fact that a lot of hesitant responses probably occur. As a result, as indicated by Figure 4.7, the geographical profiles for these concepts are probably more fragmented than expected. For salient concepts, however, it was argued that even if they are more vague, there may still be one or a few lexical variants that take up a central position in their profile (recall that, for instance, for SLIM (ZIJN), one lexical item, *slim*, occurs in 48.97% of all the observations for the concept and additionally, 69 other lexemes are used as well). As the calculation of geographical fragmentation takes into account the geographical profile (i.e. the relative frequency of the single lexemes that occur for a concept), this results in an unexpectedly high number of unique types and, thus, a stronger positive correlation with lexical non-uniqueness.

4.5 DISCUSSION

The first part of the analyses indicate that the effect of the concept features remains stable for the different aspects of the response variable (the number of unique types, the weighted average dispersion and the weighted average lack of spread per concept). Although the relative ranking of the predictors in the models is comparable, for each of the three aspects of dialectal lexical diversity, it differs to some extent. Crucially, again we do not find any significant interaction effects with 'dictionary' which indicates that the results obtained are stable across dialect areas.

This part of the analysis also showed that some of the findings of Pickl need to be attenuated. More specifically, only for onomasiological vagueness, the findings were as expected: vagueness predominantly affects the number of unique types and weighted average dispersion, but takes up a less strong position for the weighted average lack of spread. These results may be explained by the fact that, as Pickl argues, demarcational choices between speakers from different locations may occur. For affect, Pickl's hypothesis (innovation affinity) was not entirely confirmed: affect only has a small effect on the weighted average lack of spread of a concept. Perhaps this has to do with the fact that lexical items that are used for affected concepts can still have a high disposition to quickly spread in space, because they are communicatively relevant. As they quickly lose their non-connoted meaning, people from different locations may quickly borrow them from each other, leading to a less strong correlation between affect and weighted average lack of spread. Finally, the results for the local versus supralocal semantic fields (diffusion affinity) was also not confirmed, as semantic field only has a small effect on each aspect of lexical diversity.

Furthermore, the second part of the analyses, which predominantly serves an exploratory purpose to examine whether the concept features are only significant *because* the data are geographically stratified, indicates that most of the concept features remain significant. The model is similar to the one that was used in chapter 3, although a separate variable selection procedure was used: onomasiological vagueness and proneness to affect correlate positively with the number of unique types, given the geographically stratified profiles of the concepts. As the type of geographical variation accounted for in this model is characteristic

of dialectal data, the second part of the analysis serves as a first indication that concept features may also affect varieties stratified along a different dimension.

Additionally, the two parts of the analyses also seem to reveal that the impact of onomasiological salience is of a different kind. On the one hand, onomasiological salience seems to be more important for the geographical stratification of the data. On the other hand, only one predictor that gauges this concept feature reaches significance in the second part of the analysis. Consequently, this indicates that, all other things being equal, the degree of onomasiological salience of a concept is predominantly predictive of the geographical fragmentation of the concept. More specifically, highly salient concepts occur with a small degree of geographical dispersion and lack of spread and, thus, highly uniform geographical profiles, in which one or a few variants are very frequent and take up clear and relatively large geographical areas in space. Highly non-salient concepts, on the other hand, show a large degree of spatial diversification, with high values for dispersion and lack of spread. However, this does not mean that onomasiological salience cannot affect lexical diversity in differently stratified varieties. Instead, the results seem to imply that, while a higher degree of onomasiological vagueness and affect induce both more heterogeneous profiles for the concepts and, over and above geographical fragmentation, a larger amount of unique variants per concept, a higher lack of onomasiological salience only affects the former aspect of lexical diversity. From a prototype-theoretical perspective, these non-identical effects of onomasiological vagueness and affect, on the one hand, and onomasiological salience, on the other, are not surprising. For concepts that are onomasiologically less salient, be it on a low taxonomical level or between 'co-hyponymous' concepts, research has shown that hyperonymous, co-hyponymous, or possibly hyponymous names that are associated with more salient concepts are often used (Berlin, Breedlove & Raven 1973, Geeraerts, Grondelaers & Bakema 1994). For less vague and affect-sensitive concepts, however, the framework outlined in Pickl (2013) indicates that a disposition to lexical creativity or to demarcational differences between speakers, are the mechanisms that correlate with these features. These latter mechanisms cause a higher amount of unique lexical types for vague concepts (in the form of a large amount of inconsistent responses between dialect speakers) and for affect-sensitive concepts (because the dialect speakers continuously (need to) come up with innovative ways to refer to these concepts). However, because we only inquired into geographically stratified onomasiological profiles, further research is necessary to confirm this interpretation.

Crucially, then, this chapter shows that the effect of the concept features is dependent on the definition of lexical diversity. This definition often depends on the aim of the research. As outlined in the methodology section, by using our method of accounting for the geographical signal in the data, we predominantly inquire into the availability of synonymous, equivalent expressions for a particular concept. In this sense, this study is comparable to research in language evolution, in which lexical diversity (more specifically, the rate of lexical replacement) is operationalized as the number of unique lexical variants that are available per concept from a diachronic perspective (e.g. Pagel, Atkinson & Meade 2007). In research into lexical richness, however, the operationalization of lexical diversity also takes into account the number of tokens per concept, by, for instance, using type-token ratios or guiraud scores (e.g. Daller, Van Hout & Treffers-Daller, Jarvis 2013, Tweedie & Baayen 1998). However, from a Cognitive Sociolinguistics perspective, the stratification of the data, be it along a geographical, sociolinguistic, register-dependent, or other axis, should be taken into account as well. For instance, if the aim is to gauge the extent to which a particular variant is dominant in a speech community (which may be sociolinguistically, geographically or otherwise stratified), then taking an onomasiological perspective that accounts for differences in the frequency of all the variants that occur for a particular concept, is necessary (see Geeraerts, Grondelaers & Speelman 1999).

As we have merely examined a relatively naïve way of measuring lexical diversity if the geographical signal is accounted for, the results that we obtained should be complemented in two ways. First, different methods for taking into account the geographical signal in the data can be envisaged. To obtain further evidence for the validity of the different results that were obtained for onomasiological vagueness and affect, on the one hand, and onomasiological salience, on the other, the degree to which variation in the amount of lexical variants that occur shows standardization, given the geographical distribution of the data, would have to be investigated. This can be accomplished by examining the effect of the concept features on variability in the amount of lexical standardization in smaller, linguistically consistent dialect areas within the broader regions where the Brabantic and Limburgish dialects are used. Second, to determine whether the results obtained in this chapter hold in differently stratified languages, research into other varieties is necessary as well. However, one of the main advantages of the dialect data that are used throughout this dissertation is that they are part of a large-scale onomasiological database of lexical variation. Similar datasets for differently stratified data are not as easily available: to obtain onomasiological profiles in corpus data, for instance, a lot of manual disambiguation

is still necessary, which requires a lot of time and, therefore, has the disadvantage that only a particular part of the lexicon (i.c. only a few concepts, registers or semantic fields) can be taken into account (e.g. Daems, Zenner & Geeraerts 2016, Ruette, Speelman & Geeraerts 2014). For this reason, we believe that using the method outlined above is valid for serving its preliminary and exploratory aim.

Overall, then, the first part of this dissertation revealed several aspects of the degree to which concept-related features affect lexical diversity in dialectal data. Chapter 3 showed that onomasiological vagueness, onomasiological lack of salience, and proneness to affect correlate positively with the degree of lexical diversity a concept shows. Furthermore, these findings are stable across semantic fields and dictionaries, although differences in the effect size of the concept features do exist between semantic fields. The current chapter indicates that the way in which these concept features affect lexical diversity can be of different types. Onomasiological vagueness and affect influence both the degree to which a concept shows geographical fragmentation and the number of alternative variants that are available. Onomasiological salience correlates with the number of unique types to a lesser extent, given the weighted geographical stratification of the concept, although it seems to be influential for the degree of uniformity in the geographically stratified onomasiological profile of the concept.

In the next chapters, we inquire further into the effect of concept-related features on lexical diversity, but we take a different approach. More specifically, we more explicitly recognize the Cognitive (Socio)linguistic premise that meaning is experiential and encyclopaedic by taking into account the fact that concept-related characteristics need not be homogeneous throughout the dialect areas. In chapter 5, we use the geographical signal in the data as a response variable to investigate to what extent onomasiological geographical variation reflects patterns in socio-cultural history. In chapter 6, we use geographical variation in the socio-cultural environment of the dialect speaker as an operationalization of experiential salience and examine whether and how this variable correlates with lexical diversity in dialect data.

Case studies – part 2

5. Formal variation in dialect data: Semantic and geographical patterns in the distribution of loanwords

5.1. INTRODUCTION

Part 1 of this dissertation showed that cognitive concept features, like the degree of onomasiological salience, vagueness and proneness to affect of a concept, influence the amount of dialectal variability the concept shows. However, a question that remains unanswered is how this type of lexical diversity is formally reflected in the language use of a dialect speaker: we have not yet examined whether the interaction between meaning and lectal features also affects the lexical variants that are used. Do the types of lexical items that are chosen for particular concepts also reflect semantic differences? To what extent are such differences influenced by a dialect speaker's geographical location? We will inquire into these questions by analysing geographical and semantic structure in the use of loanwords from different source languages throughout the Brabantic and Limburgish dialect area.

By conducting such an analysis, we contribute to Cognitive Linguistics in two ways. On the one hand, while it is known that the use of a particular variant can elucidate the way language users structure their everyday environment, lectal (in this case: geographical) differences are not always taken into account. On the other hand, the fact that the use of loanwords is influenced by the meaning to be expressed has been discussed before as well: loanwords provide insight into the influence of cultural variation on the structure of the lexicon. These findings were already noted by Paul (1891[1880]: §698): he asserts that the use of loanwords may be restricted to particular groups, connected by social ties or characterized by geographical proximity. However, few studies take into account both a large, stratified and semantically varied dataset of naturalistic material, and different source languages at the same time. A notable exception is Geeraerts, Grondelaers & Speelman (1999) who examine variation in the use of loanwords from French and English in two varieties of Dutch.

First, attention for the lectal dimension is, for instance, sometimes missing from research on the relationship between naming and categorisation. A body of evidence indicates that, in various languages and in differently stratified varieties, the use of a particular item can, for instance, reflect the degree of prototypicality of the referent (e.g. Berlin, Breedlove & Raven 1973, Berlin & Kay 1969, Rosch 1978, Swanenberg 2000), or salient features of the referential meaning to be expressed (e.g. Lakoff & Johnson 1980, Langacker 2008: 55-89, Talmy 1985, 2000, also see the recent volume by Blumenthal-Dramé, Hanulíková & Kortmann 2017). Typological and anthropological research has also shown that whether speech communities coin a name for a particular concept or rely on descriptive and, thus, not lexicalized phrases instead, can be related to cultural differences and the experiential environment of the language users (Majid & Burenhult 2014). Nowadays, most scholars agree that categorization also depends on differences between cultures (e.g. Kövesces 2005), Nonetheless, the degree to which the use of a particular type of name is variable or systematic within a speech community, is less well researched. Recently, including the lectal dimension for systematic research on lexical borrowing has been receiving more attention (e.g. Zenner & Kristiansen (2014), in which the need for an onomasiological and usage-based approach is advocated).

Second, attention for meaning in language contact often takes the form of providing an overview of the degree of borrowability of lexical items to express particular meanings, often by referring to the lack of borrowability of core vocabulary (e.g. Hock & Joseph 1996: 257, Thomason 2001: 70-72), although the importance of culture is frequently mentioned as well (e.g. McMahon 1994: 201-204). For instance, one of the results of the Loanword Typology Project (Haspelmath & Tadmor 2009) is that the amount of loanwords that are borrowed in the world's languages, differs dramatically per semantic field. (Tadmor 2009: 64-65).

Table 5.1 shows the distribution of loanwords per semantic field across the languages included in the project. Fields that are prone to borrowing across varieties include 'religion & belief' (41.2 % loanwords), 'clothing and grooming' (38.6 % loanwords) and 'the house' (37.2 % loanwords), while only very few instances of lexical borrowing occur in the field of 'sense perception' (11 % loanwords), 'spatial relations' (14% loanwords) and 'the body' (14.2% loanwords). According to Tadmor, this has to do with the fact that the former fields are more heavily influenced by cultural interactions, while the latter contain concepts that are universal. For this reason, each language contains native elements to express these concepts. However, projects like this do not pay attention to the degree to which, within a single language, lectal features may be relevant as well. A notable exception is Zenner, Speelman & Geeraerts (2014).

Additionally, due to the nature of the dataset used, one practical advantage is important as well. More specifically, in the datasets of the WBD and WLD, the amount of loanwords that are used, can be calculated automatically, which allows for a large-scale, systematic investigation into patterns in the geographical and semantic systematicity in loanword usage. More specifically, in these databases, a loanword is marked with a specific tag, like 'fr.', 'du.' or 'lat.', which also reveals the source language of the borrowed item. As a result, the amount of variation in the use of this naming strategy can be calculated automatically and a very large amount of data can be taken into account at the same time. Furthermore, as the concepts that are examined contain data from the entire dialect areas, the onomasiological perspective is safeguarded. Consequently, we follow the Cognitive Contact Linguistics perspective on borrowing which advocates the importance of multifactorial, large-scale and mixed-data approaches (Zenner & Kristiansen 2014: 10). However, whereas Zenner & Kristiansen propose this perspective as a means to acquire insight into the process of lexical borrowing (ibid.: 1, 5), this chapter takes a different perspective: what does lectal and semantic variation in lexical borrowing reveal about the way the lexicon of the receptor varieties are structured?

Additionally, this big data approach ties in with recent advances in lectometry. In contrast with traditional dialectological research, which comprised a certain amount of subjectivity due to the fact that dialectologists were unable to aggregate over a large amount of features and locations at once, dialectometrists argue that "computational and statistical analysis now makes it possible comprehensively to compare inventories attributively drawn from a great many locations" (Nerbonne & Kretzschmar 2003: 4-5). By extending the scope to data stratified in other ways, recent studies have demonstrated that using large data sets and quantitative methodologies not only allows for the identification of variation of a geographical nature (e.g. De Vriend, Swanenberg & Van Hout 2007, Goeble 2006, 2010, Grieve 2013, Grieve, Speelman & Geeraerts 2011, Heeringa 2004, Heeringa & Nerbonne 2006, Nerbonne & Kleiweg 2003, Pröll 2013, Séguy 1971, Szmrecsanyi 2013, Wieling, Upton &

semantic field	loanwords as % of total
religion and belief	41.2%
clothing and grooming	38.6%
the house	37.2%
law	34.3%
social and political relations	31.0%
agriculture and vegetation	30.0%
food and drink	29.3%
warfare and hunting	27.9%
possession	27.1%
animals	25.5%
cognition	24.2%
basic actions and technology	23.8%
time	23.2%
speech and language	22.3%
quantity	20.5%
emotions and values	19.9%
the physical world	19.8%
motion	17.3%
kinship	15.0%
the body	14.2%
spatial relations	14.0%
sense perception	11.0%
all words	24.2%

TABLE 5.1
Borrowing per semantic field in the Loanword Typology Project (Tadmor 2009: 64)

Thompson 2014), but also facilitates distinguishing patterns structured along a social, discursive or cultural axis (e.g. De Pascale Forthcoming, Grieve 2007, 2017, Heylen et al. 2015, Luyckx & Daelemans 2011, Ruette, Ehret & Szmrecsanyi 2016, Ruette & Speelman 2013, Speelman & Heylen 2017, Wieling 2012, Wieling, Nerbonne & Baayen 2011). In this paper, we employ comparable dialectometric techniques to inquire into the interplay between geography and meaning.

In sum, the combination of the theoretical advantages of loanword research (viz. insight into cultural and geographical variation) with the practical benefit of using a large, automatically collected dataset, will allow for a large-scale investigation into the extent to which loanwords from different source languages are distributed systematically across the dialect areas under scrutiny. Section 5.2.1 provides an overview of the relationship between the Brabantic and Limburgish dialect areas and other languages and varieties, and of the cultural patterns that have been shown to be relevant for the use of loanwords in Standard Dutch. In 5.2.2, the hypotheses that can be distinguished on the basis of these patterns are outlined. Section 5.3 describes the data and methodology used in this chapter. In section 5.4, the results of the analyses are presented, followed by a discussion in section 5.5.

5.2. LEXICAL BORROWING IN THE DUTCH LANGUAGE AREA

5.2.1 Geography and cultural history: differences between source languages

Two processes can be distinguished that influence variation in the use of non-native lexical items along the geographical dimension. First, language contact between languages that are geographically nearby can cause interference between the varieties. These languages may or may not be closely related (cf. Weinreich 1968: 1-2). Second, throughout western European history, several diglossic constellations have been relevant, in which a particular exoglossic standard exerted its influence on everyday language (Auer 2005). Such diglossic constellations are heavily dependent on the geography of historical evolutions: not every region has been influenced to the same extent by the same languages. The Brabantic and Limburgish dialects are particularly interesting, because, as a result of their geographical position and due to the socio-political history of the Dutch language area, they are susceptible to a complex constellation of different types of language contact, which results in borrowing from different source languages. The Dutch language area has throughout history been influenced by contact with several cultures and languages, including French, Latin, English and, to a lesser extent, German.

Additionally, for standard Dutch, several studies in historical sociolinguistics have shown that the influence of exoglossic varieties is often especially dominant for particular semantic fields of the receptor language. The influence of French is investigated in Salverda de Grave (1920) and, more recently, a collection of papers concerning the influence of the language border between the Germanic and Romance languages has appeared (Peersman, Rutten & Vosters 2015). German borrowings in Dutch are discussed in Theissen (1975). The distribution of Latin in Dutch is researched in Van der Sijs & Engelsman (2000) and Weijnen (1967). These and other studies are summarized in Van der Sijs (2005). Recently, research in Cognitive Sociolinguistics and variationist linguistics has inquired further into the geographical and semantic distribution of loanwords in colloquial or dialectal varieties of Dutch from an onomasiological perspective. Examples include Daems, Heylen & Geeraerts (2015), Geeraerts et al. (1999), Van Hout, Kruijsen & Gerritsen (2014) and Zenner (2013). This chapter complements these studies by taking into account lexical borrowings from more than one source language in several semantic fields at the same time in dialectal varieties.

French

According to Van der Sijs (2005), due to a long history of contact with the French people, French loanwords are, overall, widely accepted in Dutch and occur in a variety of semantic fields, including military (e.g. *artillerie* 'artillery', *luitenant* 'lieutenant'), the arts (e.g. *melodie* 'melody' and *gravure* 'engraving') and everyday language (e.g. *fauteuil* 'armchair' and *blouse* 'shirt'). Furthermore, French loanwords are frequently used in Standard Dutch for concepts relating to administration and government, which can be explained by the fact that French administration and law were introduced in the Low Countries during the Napoleontic regime (1795-1813). Additionally, French was used for these purposes even longer in the northern part of Belgium, until the Flemish movement gained political ground and Dutch became the official language of politics, education and administration in the 1930s. Daems et al. (2015) and Geeraerts at al. (1999) inquire into the use of French for a semantic field pertaining to everyday language, viz. clothing concepts (also see Van der Sijs 2005: 184). They clearly find diverging patterns between Belgium and the Netherlands. More specifically, due to its complex relationship with the French culture, Belgian Dutch seems to react, in a purist fashion, against the abundance of French loanwords in the language, which is apparent from

the decreasing number of French loanwords between the 1950s and 2012 in the field of clothing terminology. A similar reactionist tendency is absent in Netherlandic Dutch.

Additionally, the Germanic-Romance border is found in the south of the Brabantic and Limburgish dialect area. Research on lexical borrowing has indicated that the distance hypothesis (i.e. the further away from the border, the smaller the amount of borrowing from French) holds in the local dialects of Dutch located near this border (Kruijsen 1990). However, as the language border only became a political border in the 1960s, the amount of French items used by a speaker is also dependent on their age and on the amount of contact they have with francophones (Van Hout et al. 2014). Additionally, in the south of the Brabantic dialect area, the city of Brussels is located. In this city (and in the surrounding towns), French has always played an important role. Initially, it only served as the language of the nobility, but due to the fact that the French language was much more prestigious than the local Brabantic dialects, the number of people who used a variety of Dutch decayed over time, in favour of the French language (De Vriendt 2004: 20-29 and 91-94).

Latin

Latin has exerted its influence on the Dutch language in various domains and throughout time. Van der Sijs & Engelsman (2000) mention the influence of Latin on the Germanic languages during the Roman era in semantic fields like military and politics (e.g. *defensie* 'military defense' and *pijl* 'arrow'), trade (e.g. *munt* 'coin' and *kopen* 'to buy') and the names for days of the week and for the months. In medieval times, Latin was mostly important as the language of the Catholic church but it also exerted its influence on Dutch for concepts relating to education (e.g. *school* 'school', *schrijven* 'to write'), science (*epidemie* 'epidemic', *recept* 'recipee'), and for administration and government (*artikel* 'article', *decreet* 'decree'). Furthermore, words from Church Latin were borrowed for novel religious concepts when the people of the Low Countries were christened (Van der Sijs 2005: 124). Semantic fields that were influenced by Latin during the Renaissance period, finally, include the field of higher education (e.g. *academie* 'academy', *docent* 'university teacher') and administration and government (e.g. *agenda* 'calendar', *collega* 'colleague'). Crucially, in comparison to French, the use of Latin is probably less prone to geographical variability, as Latin has predominantly been influential as a written, academic language. Political conflicts between the Germanic tribes and Roman people, who spoke a variety of Latin as their native tongue, only occurred in the Roman era.

German

In the east of the province of Limburg, a border with Germany is found, which was installed at the beginning of the 19th century. This border is interesting as German and Dutch are closely related West-Germanic languages. More specifically, the Germanic and Dutch dialects historically form a continuum: some of the dialects spoken in the south of Limburg in the Netherlands can even be considered dialects of German, as they underwent the second Germanic consonant shift (viz. the Ripuarian dialects, see Van de Wijngaard & Keulen 2007). Research into the effect of the border with Germany in the Kleverland dialect continuum in the north of Netherlandic Limburg has shown that it has come to serve as a social and linguistic boundary and that the dialects on each side of the border show signs of convergence with their respective standard varieties (Giesbers 2008, De Vriend et al. 2014). The extent to which the Ripuarian dialects have been influenced by the language border has been less systematically researched (but see Cornelissen 2007 for an overview of relatively recent loanwords from German and Roukens (1961) for a brief history of the dialect of Kerkrade).

According to Van der Sijs (2005: 257-259; also see Weinreich 1968: 1-2), the fact that German and Dutch are closely related results in a smaller amount of loanwords that are clearly German in the Dutch Standard language, because they are often borrowed in a "dutchified" form (e.g. *bespreken* 'to dicuss' (Germ.: *besprechen*), *drukknoop* 'press-stud' (Germ.: *Druckknopf*) and *warenhuis* 'department store' (Germ.: *Warenhaus*)). Furthermore, although the German language area shares a border with the region where Dutch is spoken, the influence of French has always been greater, because of the great importance of French culture throughout Europe since the Middle Ages (Van der Sijs 2005: 268). In Standard Dutch, the semantic fields in which the influence of the German language and culture are clear, are trade, religion, science and warfare (Van der Sijs: 274-286). Trade terminology was predominantly borrowed through trade contacts with the Hanse in the Middle Ages. As a result, the Dutch language contains Middle Low German words like *eigenwijs* 'precocious', *daalder* 'thaler' and *kroeg* 'pub'. After the Middle Ages, High German became the dominant variety. Many religious loanwords stem from after the Reformation, when the Luther Bible was translated from (High) German into Dutch, like *afvallig* 'unfaithful', *heftig* 'fierce, intense', *slachtoffer* 'victim'. In the 19th century, German culture was influential in areas like science (e.g. *bewusteloos* 'unconscious', *psychoanalyse* 'psychoanalysis' and *volksetymology* 'folk etymology'), socialism and politics (e.g. *jeugdbeweging* 'youth movement', *kartel* 'cartel' and *autobaan* 'motorway') and industry (*erts* 'ore', *benzine* 'petrol' and Fahrenheit).

Finally, German words in Dutch having to do with warfare and army are *schermutselen* 'to skirmish', *hamsteren* 'to hoard' and *concentratiekamp* 'concentration camp'.

Other languages

Loanwords from other languages occur in the Dutch language as well. For instance, Zenner, Speelman & Geeraerts (2012) show that, recently, lexical borrowing from English has become especially frequent in the semantic fields of media and IT. However, as we focus on concepts concerning the everyday life in the traditional agrarian society, loanwords from English are scarce. Borrowings from other languages are very infrequent as well: the dataset that will be used in the remainder of this chapter only contains 16 non-native word types borrowed from other languages (viz. 11 from Italian, 2 from Greek (in the field of church & religion) and 1 Portuguese, Spanish and Hungarian loanword). For this reason, the remainder of this chapter will focus on the distribution of loans from French, Latin and German.

5.2.2 Hypotheses

The previous section showed that variation in the use of loanwords will probably be influenced by the interaction between socio-cultural history and the geographical location of a dialect speaker. To investigate this interaction in the Brabantic and Limburgish dialects, we focus on four volumes (i.e. semantic fields) of the digitized dictionaries:

- III.1.3: clothing & personal hygiene
- III.1.4: personality & feelings
- III.3.1: society, school & education
- III.3.3: church & religion

As will be discussed in more detail below, these semantic fields were chosen because they are expected to show clear patterns of geographical and cultural variation in the use of loanwords. Furthermore, while most of the fields are prone to borrowing according to Tadmor (2009), the field of personality & feelings takes up a special place.[1]

As outlined above, concepts from the semantic field of clothing & personal hygiene, are part of the everyday language of a dialect speaker and are prone to lexical borrowing (38.6% borrowed items in Tadmor 2009). Additionally, detailed research into this field has shown that the use of French loanwords is especially frequent, although clear differences between Belgian and Netherlandic Dutch occur (Daems et al. 2015 and Geeraerts et al. 1999). As the dialectal data we use come from an early time period (87.5% of the clothing data in the database were collected in the 1960s) and from a different variety (viz. from the base dialects of Dutch), our data serve as a historical, differently stratified alternative to the oldest data used in Geeraerts et al. (1999).

In the semantic field of society, school & education, lexical borrowings from both French and German are expected. Table 5.2 shows the subdomains in this field in the Dictionary of Limburgish Dialects.[2] On the one hand, it contains concepts relating to the military, politics and education, which have been argued to often be expressed with French items. Additionally, as French culture was dominant for a longer period in the northern part of Belgium in this field, we expect to find differences between Belgium and the Netherlands. On the other hand, trade and industry concepts, which can be related to German culture, are included as well. In as far as this field is included in Tadmor's division, it is expected to show a relatively high amount of loanwords as well (law: 34.3% loanwords, social and political relations: 31.0% loanwords, warfare and hunting: 27.9% loanwords, possession: 27.1% loanwords, speech and language: 22.3% loanwords).

The field of church & religion is chosen because, according to Tadmor (2009), this field is highly susceptible to borrowing. The use of Latin lexical borrowings is expected to be especially frequent in this field, although some German loanwords may be used as well. However, we expect to find no geographical variation for the distribution of loanwords in this field, as the concepts in the database refer to practices in the Catholic church, which were frequent throughout Limburg and Brabant (Schmeets 2014). Additionally, many of the church-related Latin words were introduced as names for novel concepts.

Finally, we also include the semantic field of personality & feelings. Table 5.3 shows the subdomains of this semantic field in the WLD. On the one hand, this table contains some concepts, relating to feelings/emotions and values, that are not prone to borrowing according to Tadmor (2009, see Table 5.1), because they contain universal/core vocabulary concepts. On the other hand, some subsections of this semantic field, like behavioural traits or affect-sensitive concepts (e.g. concepts related to indecency or stupidity), may also require a certain degree of personal involvement. As a result, if we do find a large number of loanwords in this field, perhaps this has to do with the "need for synonyms" of the speakers, which allows them to retain the expressive force of affected concepts (Weinreich 1968: 58-59 and cf. part 1). However, if we do not find that loanwords are used

1 It should be noted that the semantic fields distinguished by Haspelmath & Tadmor are not identical to the semantic fields in the WLD and WBD, nor do they contain exactly the same concepts. However, as both were collected on a very large scale, we assume that the general patterns are comparable.

2 The subdivision into subdomains is almost identical in the WLD and WBD across dictionary volumes.

subsection	examples
man and society	e.g. trade, money, property, labour, language, communication
societal organisation	e.g. societal institutions, taxes, elections, police, law and crime, defence and war
transportation	by road, by railway, by air, over water
education	e.g. people in school, the school building

TABLE 5.2
Subdomains of the field of society, school & education in the WLD

subsection	examples
intellectual capacity and memory	e.g. thinking, knowing, smart, dumb, to judge/to consider
personality	e.g. (un)reliable, (in)sincere, diligent-lazy, brave-frightened, conceited(ness)
feelings	e.g. fun, laughter, anger, sadness, disappointment
behaviour	e.g. to behave, dominance, to (dis)obey, success-failure, (in)decency

TABLE 5.3
Subsections of the field of personality & feelings in the WLD

	society, school & education	clothing & personal hygiene	church & religion	personality & feelings
expected source language(s)	French German	French	Latin (German)	no borrowing
expected geographical pattern	French: Belgium vs. Netherlands	Belgium vs. Netherlands	no geographical variation	no geographical variation

TABLE 5.4
Overview of hypotheses

source language	society, school & education		clothing & personal hygiene	
French	coupon portefeuille	‘ticket (transport)’ ‘wallet’	bijou winterpaletot	‘jewel’ ‘warm coat’
Latin	statie tribunal	‘station’ ‘cantonal court’	stola stool	‘stole’ ‘bonnet of the “poffer”’
German	rad flik	‘bike’ ‘police officer’	absatz smuk	‘shoe heel’ ‘ornament’
source language	**church & religion**		**personality & feelings**	
French	medaille voile	‘scapular’ ‘headdress for girls during Holy Communion’	bleu caractère	‘shy’ ‘personality’
Latin	crucifix monstrans	‘crucifix’ ‘monstrance’	permitteren pretentie	‘to rant and rave’ ‘pride’
German	bleien venster dirigent	‘leaded window’ ‘choirmaster’	juxig geschaft	‘comical’ ‘artificial, forced’

TABLE 5.5
Examples of loanwords per source language and semantic field

for personality & feelings concepts, we can provide further evidence for the stability of universal vocabulary.

In sum, a complex network of semantic and geographical features is expected to influence variation in the use of French, Latin and German loanwords in the Brabantic and Limburgish dialects of Dutch. First, border effects are expected to show up near the border with Germany and near the Germanic-Romance language border in every semantic field. Furthermore, French is also expected to be more frequent around the city of Brussels, where it holds a stronger position than in the rest of the language area.

Second, cultural contact will probably show up as well, through the influence of an exoglossic standard on particular semantic fields. Such an effect is predominantly expected for French, especially for concepts relating to society, school & education and clothing & personal hygiene, and for Latin, in the field of church & religion, although German items may show up for the latter field as well. However, differences in the geographical distribution between these exoglossic standards can also show up. In the use of Latin and German for church concepts, no geographical patterns are expected, as most of these loanwords were probably introduced as names for novel, institutionalized concepts, both in standard Dutch and in the base dialects. For French, in contrast, we do expect geographical differences. French culture held a stronger position in Belgium than in the Netherlands, which may result in geographical variation between the countries. These differences will probably be especially relevant in the semantic field of society, school & education, which contains institutionalized concepts relating to administration and politics. On the other hand, the importance of French culture on everyday life will most likely show up in the field of clothing & personal hygiene: we also expect to find differences between Belgium and the Netherlands here. The hypotheses are summarized per semantic field and per source language in Table 5.4.

	WBD	WLD
French	16443 (0.051)	13015 (0.059)
not French	305848 (0.949)	208353 (0.941)
Latin	4361 (0.014)	5810 (0.026)
not Latin	317930 (0.986)	215558 (0.974)
German	318 (0.001)	2317 (0.010)
not German	321973 (0.999)	219051 (0.990)

TABLE 5.6
Absolute and relative number of French, Latin and German tokens per dictionary

5.3. DATA AND METHODOLOGY

5.3.1 Measuring the amount of loanwords per location

The databases contain tags which were added manually by the lexicographers, to indicate whether a particular lexeme for a concept has a non-native origin.[3] For instance, the word *frech* for the concept BITS 'snappy' is marked as German, while the lexical item *diligence* for the concept POSTKOETS 'stage-coach, type of carriage' has a French origin. However, words that were marked as non-native in the Limburgish data were not always given the same tag in the Brabantic data and vice-versa. For example, the lexical variant *zich ambeteren* for ZICH VERVELEN 'to be bored' is marked as French in the Limburgish data, while it does not have any tag in the WBD. To ensure maximal comparability between the dictionaries, we used an automatic tagging procedure to ensure that every word that is labelled as French, Latin or German in one dictionary, has the same tag in the other dictionary. More specifically, for each source language and for each dictionary, we first made a list containing all the lexical types with a tag for this language. Then we marked all the lexical items in the full dataset that occur on this list as French, Latin or German, respectively. Table 5.5 contains example loanwords from every semantic field for every source language.

We use the loanword tags in the dictionaries to automatically collect the number of native and non-native French, German and Latin tokens per location and per semantic field. For instance, when focussing on the French terms, the Latin and German lexical items are considered as native (i.e. non-French). The same procedure is used for the other source languages. Table 5.6 provides an overview of the total number of French, Latin and German tokens in the semantic fields that were included per dictionary, and the proportion of these non-native tokens per dictionary. Clearly, overall, the proportion of French is much higher than the proportion of Latin and German. Interestingly, the proportion of French is also almost the same in the Brabantic and Limburgish data, while both Latin and German occur more frequently in Limburg.

3 We consider lexical items marked as Picardic, Old French or Walloon as loanwords from French and lexemes marked as Ripuarian as loanwords from German. Overall, only 14 types (100 tokens) from Ripuarian, 13 from Old French (261 tokens), 2 from Walloon (6 tokens) and 1 from Picardic (1 token) occur. Two lexical items, *proces* for PROCES-VERBAAL 'report of an offence' (43 tokens) and *tribubaal* for KANTONGERECHT 'cantonal court' (20 tokens), are marked as French and Latin. We considered these word types as Latin loanwords, as they are both borrowed from Latin via French (Philippa et al. 2003-2009).

In contrast with the analyses in part 1, we do not calculate the proportion of loanwords vis-à-vis other lexical variants per concept directly. Instead, the onomasiological perspective is safeguarded in the analyses because the databases contain data for a set of concepts per semantic field, for which lexical variants were elicited throughout each dialect area. Additionally, for most of the locations, we only have one or two observations per concept at our disposal (see Figure 5.1: mean = 1.417, sd = 0.561). Only two large cities have, on average, more than 5 observations per concept, namely Maastricht (mean = 8.987, sd = 5.880) and Tilburg (mean = 5.053, sd = 4.935).

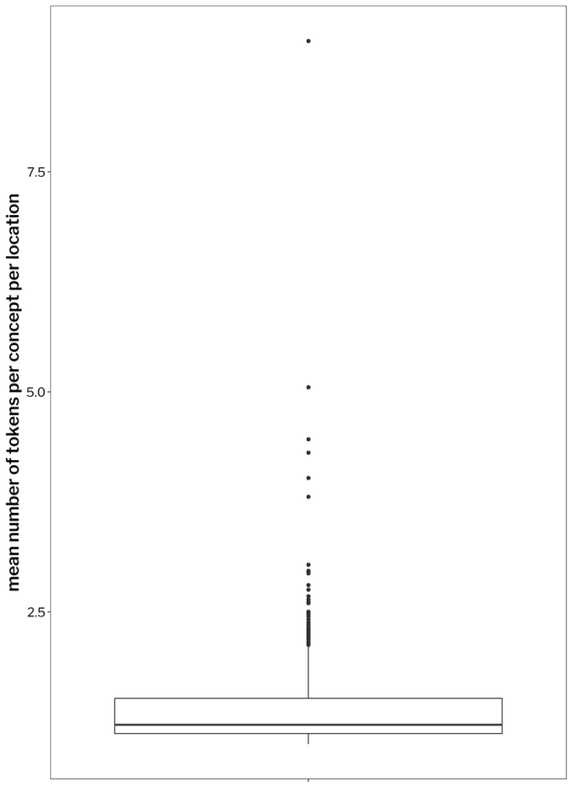


FIGURE 5.1
Boxplot for mean number of tokens per location

5.3.2 Generalized Additive Modelling

To measure the effect of the interaction between semantic field and geography on variation in the amount of French, Latin or German per location, we use Generalized Additive Mixed Modelling (GAMM). These models can be considered an extension of Generalized Linear Models. They allow for a combination of parametric and non-parametric relationships, which do not have to be specified *a priori*, between the response and the explanatory variables (Wood 2006, see Zuur et al. 2009 and Crawley 2007, chapter 18 for an accessible introduction). More specifically, they employ non-parametric smooth functions on specified model terms as part of the model fitting procedure. The models we discuss below use thin-plate regression splines to represent these smooth terms. The amount of smoothing depends on a type of cross-validation, which in practice entails that the model finds the optimal amount of smoothing while avoiding badness of fit.

With R packages 'mgcv' and 'itsadug', we build one model per source language. For each source language, we start from the same model to compare the influence of the interaction between geography and semantic field on the ratio of non-native to native tokens. This model contains a smooth term for the interaction between longitude (lon)[4] and latitude (lat) for each semantic field and a random intercept for location, as the total number of observations differs per location (although this factor does not reach significance in the model for the Latin variants). The model formula is as follows:

number of French/Latin/German tokens relative to the number of native tokens ~
semantic field + s(lon, lat, by = semantic field) +
s(location, bs = "re"), family = binomial

In our model fitting procedure, we follow the suggestions of Crawley (2007: chapter 19), Van Rij (2015), Wieling (2017) and Wood (2006: 221 - 233) and outlined in the mgcv vignette (Wood 2017). We compare AIC values and use significance tests to check whether all the predictor variables, interaction effects and smooth terms contribute to the explanatory power of the models. Finally, we visualize the predicted and the fitted values, and the residuals to assess the fit of the model to the data.

5.4. RESULTS

Figures 5.2a-c show an overview of the distribution of the proportion of borrowed lexical items in the raw data per source language in the form of bubble plots. The size of the black symbols is proportionate to the variable under scrutiny (in this case the proportion of French/Latin/German tokens per location). Black dots indicate that one or more loanwords were found. The larger the black dot, the more loanwords occur in that location. If a red symbol is present, this means that, while data for this location is available in the dictionaries, it does not contain any non-native tokens.

4 We collected longitude and latitude information semi-automatically, using the Google Maps API (see https://www.r-bloggers.com/using-google-maps-api-and-r/, Accessed on 3 July 2017).

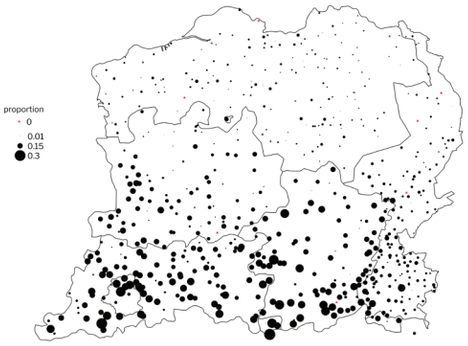


FIGURE 5.2A
Proportion of French tokens per location

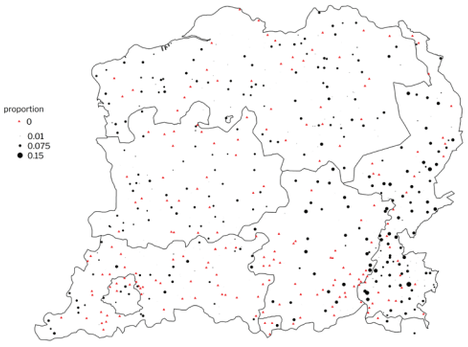


FIGURE 5.2B
Proportion of Latin tokens per location

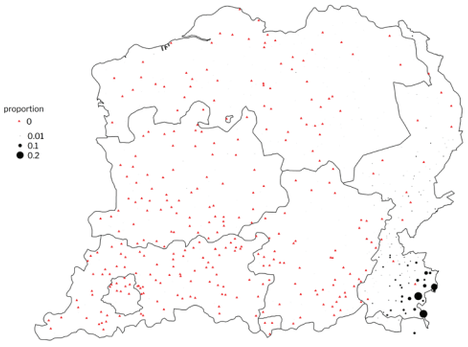


FIGURE 5.2C
Proportion of German tokens per location

Figures 5.2a-c indicate that, overall, French loanwords are much more frequent than lexical items from the other source languages. Furthermore, the figures seem to show clear geographical patterns. French is more frequent near the language border with Wallonia in Belgium, where French is spoken, and in the Dutch-speaking part of Belgium in general. German occurs the most near the border with Germany. Loanwords from this language are especially frequent in three locations near the language border (viz. in Simpelveld, Vaals and Kerkrade, which all belong to the Ripuarian dialect area). Unexpectedly, the distribution of the Latin tokens does show a geographical pattern: they seem to occur the most in the Limburgish provinces, especially in the Netherlands. The following sections aim to explain the variation in these bubble plots.

5.4.1 French loanwords

The general picture

Using the formula described above, with an interaction between longitude and latitude by semantic field and a random effect for location, we construct a Generalized Additive Mixed Model. Table 5.7 shows the relevant numerical output of the GAMM for the amount of French per location. The table should be read as follows. The upper part (parametric coefficients) provides the general effect of semantic field (the surfaces are centered in this output). In the first column, the semantic field is listed (the semantic field 'society, school & education' is included in the intercept). The second and third column show the estimate for each semantic field in comparison to the intercept and the corresponding standard error. The final column contains the p-value for each estimate. On average, we find a significantly smaller amount of French in the semantic fields of personality & feelings and church & religion in comparison to the field of society, school & education, while a significantly larger proportion of French tokens occurs in the field of clothing & hygiene.

The second part of the table (approximate significance of smooth terms) contains information about the smooth terms. The p-values in this part of the table can be used to determine whether the smooth is significantly different from 0. These values are approximate, but since they are all smaller than 0.01, they can be trusted (Zuur et al. 2009: 67). The interpretation of the smooth terms can only be done visually, although the edf (Effective Degrees of Freedom) indicate how much smoothing was used for the model term (Zuur et al. 2009: 52-53). Both the interaction between longitude, latitude in each semantic field and the random effect for location contribute significantly to explaining the variation that we find. The model particularly finds non-linear patterns for the random effect for location and in the semantic fields of society, school & education and clothing & personal hygiene.

The final part of the table (explanatory power) shows two measures that describe how much of the variance in the data is explained by the model. Overall, the model performs well. It explains 92% of the null deviance. Adjusted R^2, a value that ranges from 0 to 1 and that is another estimate for the

parametric coefficients			
	estimate	**SE**	**p-value**
intercept	-3.030	0.017	< 0.001
semantic field (*personality & feelings*)	-1.187	0.028	< 0.001
semantic field (*church & religion*)	-0.164	0.021	< 0.001
semantic field (*clothing & hygiene*)	0.740	0.018	< 0.001
approximate significance of smooth terms			
	edf	**p-value**	
s(lon, lat) : sem. field (*society, school & education*)	24.851	< 0.001	
s(lon, lat) : sem. field (*personality & feelings*)	18.033	< 0.001	
s(lon, lat) : sem. field (*church & religion*)	3.834	< 0.001	
s(lon, lat) : sem. field (*clothing & hygiene*)	21.961	< 0.001	
s(location)	329.285	< 0.001	
explanatory power			
null deviance explained	92%		
adjusted R^2	0.908		

TABLE 5.7
Numerical output of the GAMM for French loanwords

amount of variation explained, is high as well: 0.908. We also used diagnostic plots to verify that the assumptions of the model were met.[5]

Figure 5.3a-d shows the graphical output of the GAMM, with the predicted surface for each semantic field presented in a separate panel. In each panel, the Brabantic and Limburgish dialect areas are depicted, with province and country borders indicated in black. A continuous colour scale is plotted over this geographical area, with yellow hues indicating that the ratio of French to non-French tokens is high and red hues indicating that the amount of French tokens is lower. In areas where the predicted amount of French tokens is smaller than 0.03 (the lower bound of the continuous colour scale), the plots show no colour. The colours used in this figure, as well as in the other figures in this chapter, require some further attention. First, it is important to note that stronger, darker colour hues (the reddish ones) indicate a smaller amount of non-native tokens. Additionally, white regions can indicate that no borrowed lexemes are available; in this case, the white areas are delimited from the rest of the plot with non-smooth, discontinuous boundaries (like in Figures 5.3a, b and c). However, very bright hues of yellow may resemble white as well, although these light hues indicate that the amount of non-native tokens is very large.

5 More specifically, we verified that there is no harmful structure in the residuals (i.e. homoscedacity), that the residuals are not autocorrelated, that they are normally distributed, that there is a linear relationship between the predicted and observed values of the response variable and that a sufficient number of basis dimensions was used to construct the model.

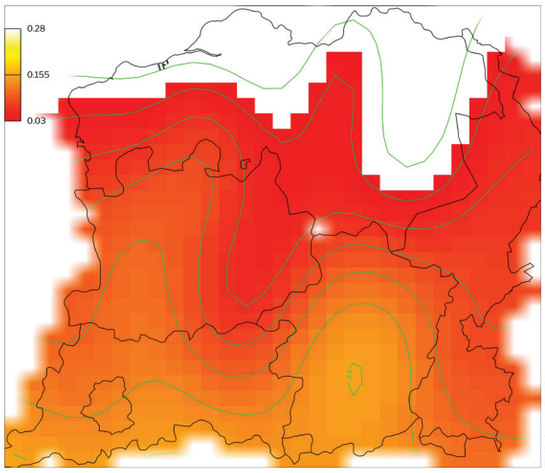


FIGURE 5.3A
Proportion of French tokens per location (society, school & education)

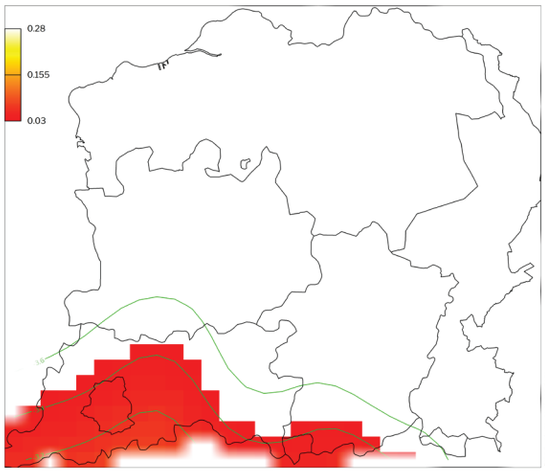


FIGURE 5.3B
Proportion of French tokens per location (personality & feelings)

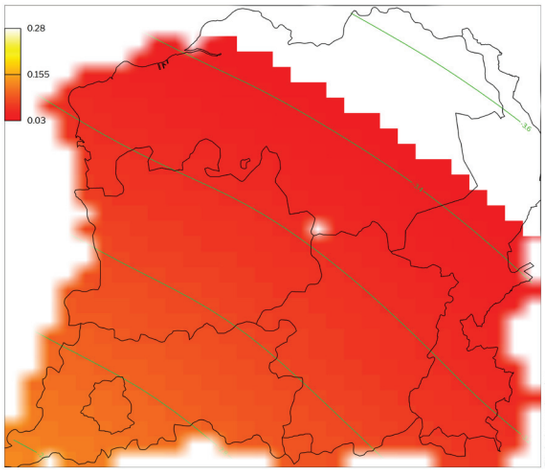


FIGURE 5.3C
Proportion of French tokens per location (church & religion)

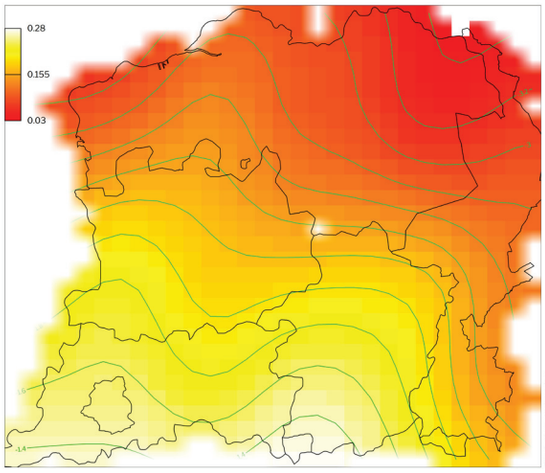


FIGURE 5.3D
Proportion of French tokens per location (clothing & personal hygiene)

Crucially, in the latter case, smooth transitions rather than discontinuous borders are shown on the plots (like at the bottom of Figure 5.3d).

The numerical interpretation of the colour scheme is provided in the legend at the top left of each panel. The legends and colour schemes are kept stable for each map per source language to ensure comparability across semantic fields (Appendix 5.1.1 contains plots per semantic field without keeping the legend stable). The minimum and maximum values for the legend are based on the predicted values for the semantic field where French occurs the most, viz. clothing & personal hygiene. Additionally, the plots also show a number of green lines that run throughout the dialect areas. These lines can be interpreted as isoglosses.

The figures confirm that, as expected, the amount of French tokens is very high in the semantic field of clothing & personal hygiene. However, French occurs even more often in this field than in the field of society, school & education. This is surprising, because military, politics and education-related terms have often been mentioned as prime candidates for the use of French loanwords. Furthermore, both maps show a clear effect of country: French is used much more frequently in Belgium than in the Netherlands. The isoglosses on the map for the clothing even seem to follow the borders of the north of the province of Antwerp and the north and the east of Limburg in Belgium. Surprisingly, however, the

lexical variant (French origin)	concept	semantic field	number of French tokens
jarretelle (fr.)	jarretelle 'suspender'	clothing & personal hygiene	161
korset (fr.)	korset 'corset'	clothing & personal hygiene	165
pelerine (fr.)	zomerkapmanteltje 'summer cape'	clothing & personal hygiene	165
jacquet (fr.)	jacquetpak 'suit'	clothing & personal hygiene	168
pitteleer (fr.)	slipjas 'tailcoat'	clothing & personal hygiene	172
voile (fr.)	rouwsluier 'widow's veil'	clothing & personal hygiene	175
corset (fr.)	korset 'corset'	clothing & personal hygiene	181
plastron (fr.)	stropdas 'tie'	clothing & personal hygiene	185
eau de cologne (fr.)	eau de cologne 'cologne'	clothing & personal hygiene	186
peignoir (fr.)	kamerjas 'dressing gown'	clothing & personal hygiene	194
pelerine (fr.)	schoudermanteltje 'shoulder cape'	clothing & personal hygiene	200
suisse (fr.)	suisse 'suisse, type of carriage'	church & religion	203
portefeuille (fr.)	portefeuille 'wallet'	society, school & education	205
caban (fr.)	wijde regenmantel zonder mouwen 'rain coat without sleeves'	clothing & personal hygiene	208
pardessus (fr.)	herenoverjas 'overcoat for men'	clothing & personal hygiene	210
soutien (fr.)	bustehouder 'brassiere'	clothing & personal hygiene	224
caban (fr.)	kapmantel 'cape'	clothing & personal hygiene	252
speculeren (fr.)	speculeren 'to speculate'	society, school & education	281
bretel (fr.)	bretel 'suspenders'	clothing & personal hygiene	285
contrefort (fr.)	hielstuk van een schoen 'heel of a shoe'	clothing & personal hygiene	294

TABLE 5.8
Top 20 most frequent French loanwords in the dataset

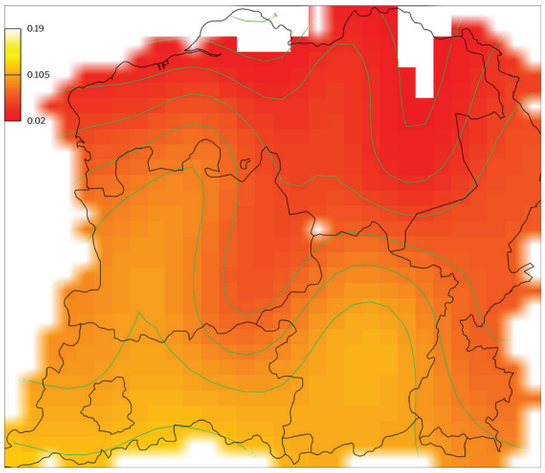


FIGURE 5.4A
Non-accepted French tokens per location (society, school & education)

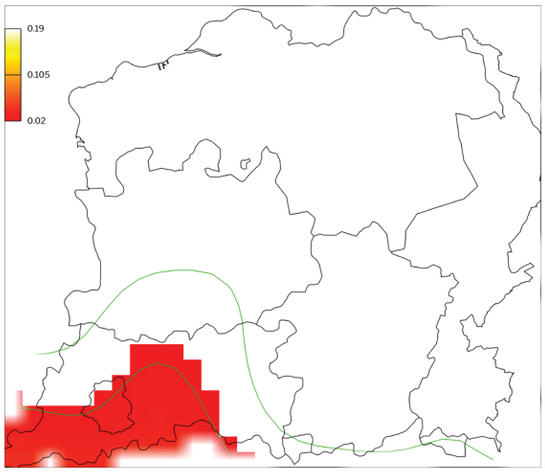


FIGURE 5.4B
Non-accepted French tokens per location (personality & feelings)

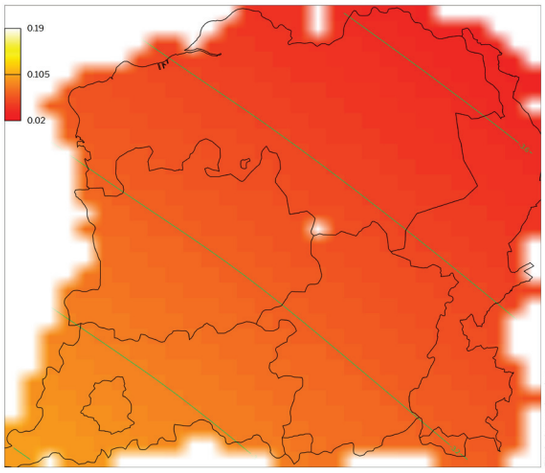


FIGURE 5.4C
Non-accepted French tokens per location (church & religion)

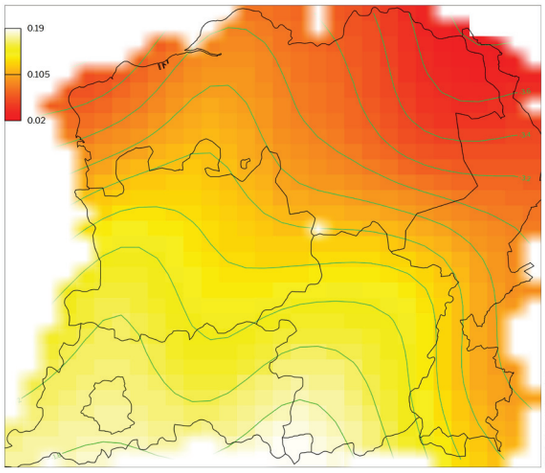


FIGURE 5.4D
Non-accepted French tokens per location (clothing & personal hygiene)

geographical patterns on both maps are not identical, as the amount of French is also relatively high in the province of Limburg in the Netherlands for society-related concepts.

French tokens are less frequent in the field of church & religion. The isoglosses, which do not show a large amount of smoothing in this field, seem to indicate that the larger the geographical distance from the city of Brussels, where French has always held a strong position, the smaller the predicted amount of French. The map for the semantic field 'personality & feelings' also shows the effect of the bilingual city of Brussels. Additionally, a language border effect seems to show up near the border with Wallonia, the Germanic-Romance border. However, overall, the amount of French is very small in this field.

The use of accepted and non-accepted French in the dialects

As French loanwords occur frequently in Standard Dutch as well, the extent to which the dialectal use of French is influenced by the standard variety does not show up in Figure 5.3. Table 5.8 shows the 20 most frequent French variants in the dataset. Many of these types are actually frequently used in Standard Dutch to express these concepts as well, which is, for instance, confirmed by the fact that the lexical variant is often identical to the name for the concept. As a result, it is unclear whether the lexical items are borrowed from French directly, or whether interference of Standard Dutch influences the geographical patterns that are found in the dialect data. More specifically, the degree to which the geographical

differences between Belgium and the Netherlands in Figure 5.3 persist if the lexical items accepted in the standard variety are excluded, is unclear.

To acquire more insight into the extent to which the dialects are influenced by the fact that some French variants are also present in Standard Dutch, we manually determined whether the available French lexemes were accepted in Standard Dutch or not at the time of the data collection process. More specifically, we rely on the 11th edition of the Van Dale dictionary (Geerts & Heestermans 1984) to verify whether each French type was already accepted in Standard Dutch at the time. We use this edition of the dictionary, because its time of publication coincides with the period in which the largest amount of data for these four semantic fields was collected (59.21% of the data was collected in the 1980s). Furthermore, while the data collection period differs between the four semantic fields (the clothing data, for instance, were mostly collected in the 1960, while data for the field of personality & feelings mostly stem from the 1980s), by the end of the 1980s, 90% of all the data in all four semantic fields was collected. If a type was not included in Van Dale 1984 ('non-accepted French'), we assume that the chance is higher that it was borrowed into the base dialects directly.

Subsequently, we conduct a second GAMM analysis to determine whether the factors distinguished in the previous model also influence variation in the amount of non-accepted French tokens vis-à-vis accepted French or native tokens (i.e. all the variants that are labelled as French in the WBD and WLD, but that do occur in Van Dale 1984, plus the variants that are not labelled as French in the WBD and WLD). We use the same model formula and verified the significance of the predictors and the assumptions of the model before interpreting the results.

Figures 5.4a-d show the visual output of this second GAMM. Interestingly, the geographical patterns in the data remain highly comparable in every semantic field. Although the overall proportion of French is smaller, as the legends show, which now run from 0.02 until 0.19 (versus from 0.03 until 0.29 in the previous maps)[6], we still find more French in the fields of clothing & personal hygiene and society, school & education. Geographically, the difference between Belgium and the Netherlands remains clearly visible in these fields as well. However, the larger frequency of French loanwords for society-related concepts in the province of Limburg in comparison to North Brabant in the Netherlands has mostly disappeared. In sum, if we take into account French loanwords that are also included in Standard Dutch, the distinction between the Limburgish and Brabantic dialect area in the Netherlands is less pronounced. Nonetheless, the border between Belgium and the Netherlands remains clearly visible. The border effect in the field of personality & feelings has been reduced to some extent: in the second model, it seems that French is predominantly used in the south of the Brussels area. Finally, regarding the field of church & religion, the figure seems to indicate that non-accepted French occurs in a larger region than in Figure 5.3. However, the map for this semantic field is slightly distorted due to the fact that the legend now has a smaller value for the lower bound (viz. 0.02 versus 0.03 in Figure 5.3 above). Overall, we can conclude that the use of French tokens in the Brabantic and Limburgish dialects is not solely due to contact with the more prestigious standard variety of Dutch.

5.4.2 Latin loanwords

To determine the influence of semantic field and geography on variation in the use of loanwords from Latin, the same model formula was used, with an interaction between longitude and latitude by semantic field and a random effect for location. However, the random effect for location does not reach significance and was therefore removed from the model. Table 5.9 shows the relevant numerical output of the GAM for the amount of Latin per location relative to the total number of tokens.

The upper part of Table 5.9 shows that, unsurprisingly, the largest amount of Latin is found in the semantic field of church & religion on average. In comparison to the reference level, society, school & education, significantly fewer Latin tokens occur in the semantic fields 'personality & feelings' and 'clothing & personal hygiene' as well.

The middle part of the table, which indicates whether the smooth terms in the model differ significantly from 0, shows that geographical variation does seem to have an effect on the number of Latin loanwords per location. In contrast with what we expected, the interaction between longitude and latitude shows significant patterns in every semantic field except 'clothing & personal hygiene'.

The bottom part of Table 5.9 indicates that the model performs very well: 94.5% of the null deviance is explained and adjusted R^2 is very high as well (0.957). We also verified the assumptions of the model. Although the model seems to struggle to a certain extent with the large differences in the amount of smoothing needed per semantic field (as indicated by the edf values in the middle part of Table 5.9), the results presented here are robust for models in which different numbers of basis functions are allowed for the calculation of the smooth term.

6 The minimum and maximum values for the legend are again based on the semantic field where French occurs the most, clothing & personal hygiene.

parametric coefficients			
	estimate	**SE**	**p-value**
intercept	-4.864	0.035	< 0.001
semantic field (*personality & feelings*)	-3.775	0.266	< 0.001
semantic field (*church & religion*)	2.239	0.037	< 0.001
semantic field (*clothing & hygiene*)	-1.636	0.084	< 0.001
approximate significance of smooth terms			
	edf		**p-value**
s(lon, lat) : sem. field (*society, school & education*)	72.126		< 0.001
s(lon, lat) : sem. field (*personality & feelings*)	15.434		< 0.001
s(lon, lat) : sem. field (*church & religion*)	16.170		< 0.001
s(lon, lat) : sem. field (*clothing & hygiene*)	16.923		< 0.1
explanatory power			
null deviance explained	94.5%		
adjusted R^2	0.957		

TABLE 5.9
Numerical output of the GAM for Latin loanwords

Figure 5.5 presents the visual output of the model for the amount of Latin per location. In these graphs, the odds of encountering a Latin token equal to 0.01 is used as the lower bound for the colour scale (in red hues). The upper bound is equal to the maximum of the predicted odds of encountering a Latin token in the field of church & religion, the field where Latin occurs the most (yellow hues). As this maximum is only 0.1, Latin loanwords are clearly less frequent overall than lexical borrowings from French (see Appendix 5.1.2 for maps of the GAM with the colour scheme differing per semantic field).

The maps for the semantic fields 'personality & feelings' and 'clothing & personal hygiene' do not contain any colour or isoglosses. This indicates that the predicted odds of encountering a Latin token in a location in these fields is even smaller than 0.01: the maximum predicted value for clothing concepts is 0.004 (in Deurne, province of Antwerp) and 0.002 for personality-related terms (in Vaals, Limburg). Recall that the numerical output of the regression model already indicated that the smooth term for clothing & personality does not differ significantly from 0.

For concepts from the field of society, school & education, Latin tokens occur in some locations, albeit very infrequently. Overall, Latin seems to be used the most for concepts of this field in the Netherlands, although the pattern seems to indicate that these tokens are geographically almost randomly distributed. Only a few locations in the north of the Belgian provinces show predicted odds between 0.01 and 0.055.

Finally, the field of church & religion shows, as expected, the largest amount of Latin tokens. As outlined above, many of the Latin names were introduced into Dutch as names for novel concepts. Whether or not a lexical item is borrowed out of necessity (i.e. to avoid a lexical gap) or

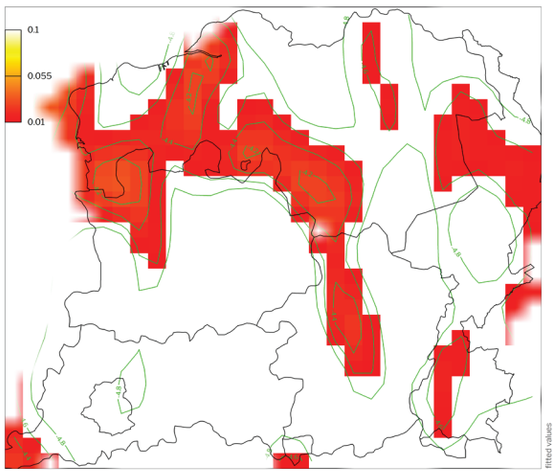


FIGURE 5.5A
Latin tokens per location
(society, school & education)

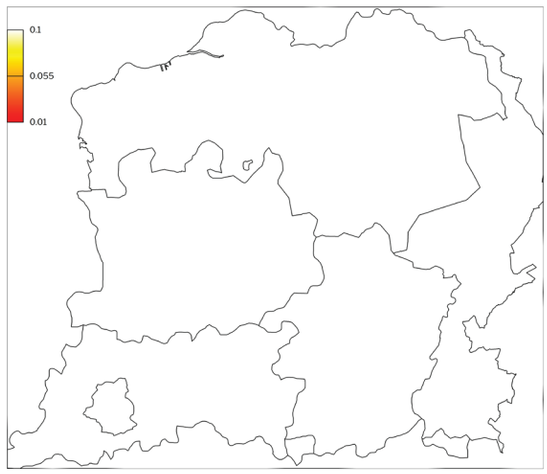


FIGURE 5.5B
Latin tokens per location
(personality & feelings)

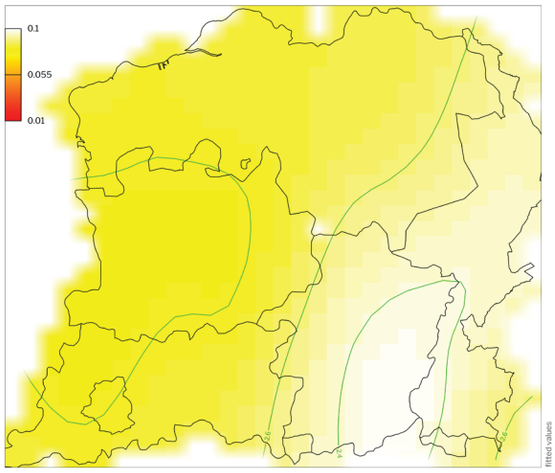


FIGURE 5.5C
Latin tokens per location
(church & religion)

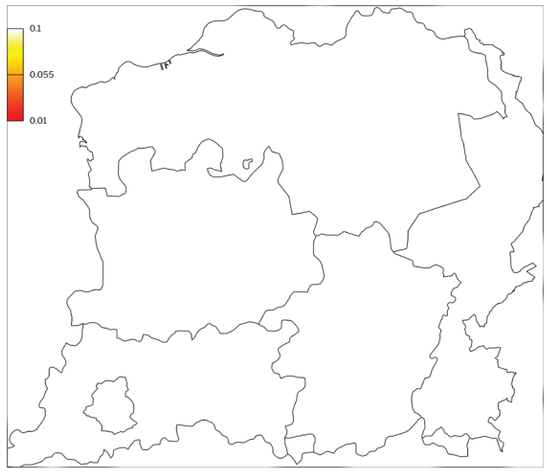


FIGURE 5.5D
Latin tokens per location
(clothing & personal hygiene)

not, is a frequently mentioned factor that increases the borrowability of a lexeme (e.g. Onysko & Winter-Froemel 2011). As a result, this factor may serve as an explanation for the success of the Latin source language in this semantic field.

However, although we did not expect to find geographical patterns in the spread of these variants for religion-related concepts, the results from the GAM indicate that Latin is used more in the two provinces of Limburg than in the Brabantic dialect area. On the one hand, the picture, thus, may reflect the boundaries between the two dictionaries that are combined in the data. It may be the case that the Latin loanwords were more consistently tagged in the WLD than in the WBD, resulting in a seemingly larger number of Latin vis-à-vis other items in Limburg. However, as outlined in the methodology section, we controlled for differences between the sources by labelling word forms that were marked in one dictionary as French, Latin or German, as foreign in the other dictionary as well (and vice versa). As a result, another interpretation seems more likely.

More specifically, this distribution can also reflect cultural patterns: perhaps people in Limburg are "more" traditionally Catholic than the Brabantic dialect users and, as a result, are more familiar with the traditional Latin names. Anecdotal evidence suggests that this interpretation holds. For instance, Limburg in the Netherlands has a higher density of pilgrimage locations that were installed in the last 200 years (Margry & Caspers 2000 :11-12). In Belgium, the percentage of catholic baptisms, weddings and funerals and of

people who attend church on Sundays in 2009 is smaller in the central cities, which are mostly located in the Brabantic dialect region, than on the countryside, which includes Limburg (Havermans & Hooge 2011). However, these differences are not completely corroborated by self-reported census data collected in the Netherlands between the 1980s and 2013. Table 5.10 shows the proportion of Catholics in the provinces of Limburg (in the Limburgish dialect area) and North Brabant (in the Brabantic dialect area) in the Netherlands, expressed in percentages (Schmeets 2014: 6). Even though today, the number of Catholics is smaller in North-Brabant than in Limburg, this diverging trend was not yet as strong in the 1980s. As the dialect data for both dictionaries in the field of church & religion were mostly collected in the late 1980s, it is unlikely that the conventionalized Latin dialect words for church concepts had already completely disappeared by this time.

year	province of North Brabant (NL)	province of Limburg (NL)
1849	100 %	100 %
1879	100 %	100 %
1909	100 %	100 %
1930	99 %	99 %
1947	98 %	99 %
1960	98 %	99 %
1971	95 %	97 %
1987	85 %	89 %
1995	80 %	87 %
1999	76 %	86 %
2003	75 %	85 %
2008	73 %	82 %
2010	69 %	80 %
2011	68 %	77 %
2012	66 %	76 %
2013	67 %	78 %

TABLE 5.10
Percentage of (self-reported) Catholics in the provinces of North Brabant and Limburg in the Netherlands from 1849 until 2013 (Schmeets 2014: 6)

5.4.3 German loanwords

For the German loanwords, we also use the same formula to determine the influence of semantic field and geography on the variation in the data, with an interaction between longitude and latitude by semantic field and a random effect for location. We verified the significance of these predictors and the assumptions of the regression model. Although the model diagnostics show that the model struggles somewhat with the general infrequency of German tokens in the dialect data, overall, it performs well. It explains 89.2% of the variation in the German versus non-German tokens (adjusted $R^2 = 0.928$, see bottom of Table 5.11).

Table 5.11 shows the relevant numerical output of the GAMM for the amount of German per location. The upper part of the table indicates that a larger number of German tokens is found in the semantic field of church & religion. The model does not find significant differences on average between the reference level, society, school & education, and the fields of personality & feelings or clothing & hygiene. The middle part of Table 5.11 indicates that all the smooth terms are significantly different from 0. Overall, the amount of smoothing does not differ much between semantic fields, although it is higher for the random intercepts for location, which indicates that there are large differences in the amount of German tokens between locations.

In Figures 5.6a-d, the predictions of the GAMM are presented visually. The minimum and maximum of the predicted values for the semantic field where German occurs the most, viz. church & religion, are used as the lower and upper bound for the colour scale in these maps, to ensure comparability between the different semantic fields. Both the minimum and maximum predicted odds are very small overall: between 0.001 and 0.113 German tokens are predicted for every non-German token. Appendix 5.1.3 contains maps for the GAMM for German variants with the colour scale chosen automatically per semantic field.

Figure 5.6 shows that the use of German tokens is geographically the most widespread in the semantic field of church & religion. While, unsurprisingly, German tokens occur in the entire Brabantic and Limburgish dialect area, they are especially frequent near the border with Germany. In the semantic field of society, school & education, the GAMM clearly shows that there is a clear difference between Belgium and the Netherlands: while no German is predicted in Belgium, it occurs much more in the Netherlands, especially near the border with Germany. Interestingly, as section 5.4.1 showed, French tokens occur more frequently in this field in Belgium. As a result, the use of German and French may be distributed complementarily: while dialect speakers from the Netherlands use German tokens, Belgian dialect users rely on French. This is not surprising as the

parametric coefficients			
	estimate	**SE**	**p-value**
intercept	-6.879	0.171	< 0.001
semantic field (*personality & feelings*)	-0.116	0.244	0.634
semantic field (*church & religion*)	0.714	0.184	< 0.001
semantic field (*clothing & hygiene*)	-0.027	0.222	0.902
approximate significance of smooth terms			
	edf		**p-value**
s(lon, lat) : sem. field (*society, school & education*)	13.905		< 0.001
s(lon, lat) : sem. field (*personality & feelings*)	11.580		< 0.001
s(lon, lat) : sem. field (*church & religion*)	14.611		< 0.001
s(lon, lat) : sem. field (*clothing & hygiene*)	12.496		< 0.001
s(location)	59.718		< 0.001
explanatory power			
null deviance explained	89.2%		
adjusted R^2	0.928		

TABLE 5.11
Numerical output of the GAMM for German loanwords

French culture held a stronger position for a longer period in Belgium than in the Netherlands. However, it should be noted that the odds of encountering a German token in the Netherlands are much lower than the odds of finding a French token in this field in Belgium, so the Netherlandic dialect speakers rely on other naming strategies for society-related concepts as well. Finally, in the semantic fields of personality & feelings and clothing & personal hygiene, a border effect shows up as well: more German is used in Limburg, near the border with Germany.

Interestingly, the border effect in three of these semantic fields (viz. society, school & education, personality & feelings and church & religion) stems from a small area in the south-east of Limburg in the Netherlands. In fact, the green lines on each of these maps, which can be interpreted like isoglosses, seem to demarcate a small area, where the use of German is exceptionally high, from the rest of Limburg in the Netherlands (also see the bubble plots in Figure 5.2). Table 5.12 shows the five locations with the highest proportion of German tokens per semantic field.[7] Notably, locations belonging to three municipalities take up a high position in every semantic field: Kerkrade, Simpelveld and Vaals. Importantly, in these locations, east of the 'Benratherlinie' (the *machen*/*maken* line), Ripuarian dialects are spoken (Van de Wijngaard & Keulen 2007, Van de Wijngaard 2007). These

7 While we included both lexical items that were marked as German and lexemes that were marked as Ripuarian in the calculation of the ratio of German tokens per location, only one token in this area is marked as Ripuarian in the data (viz. *Bohei* 'fuss', which was recorded in Kerkrade). All the other tokens that are presented in Table 5.12 are marked as High-German in the dictionaries. Furthermore, 95 out of these 100 German word types occur in the online version of the German *Duden* dictionary (http://www.duden.de, Accessed on 17 July 2017).

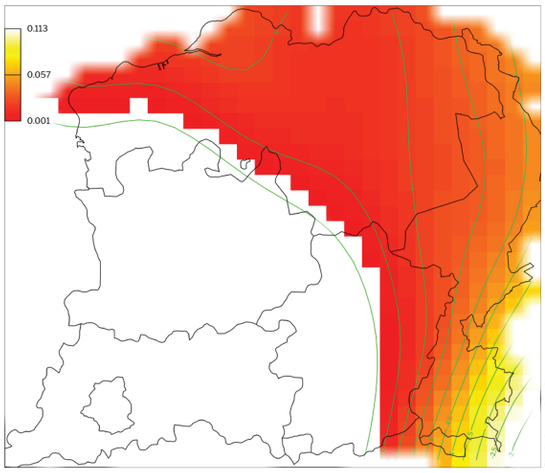


FIGURE 5.6A
German tokens per location
(society, school & education)

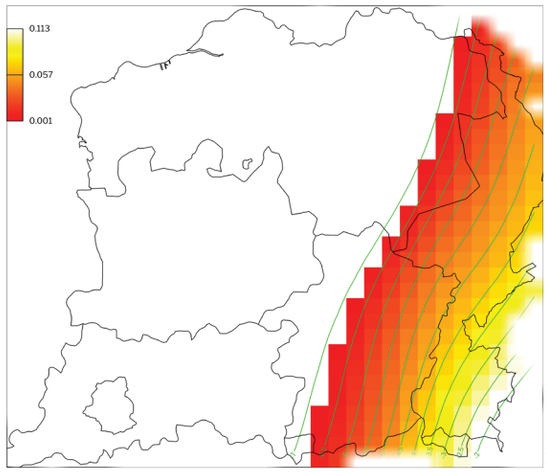


FIGURE 5.6B
German tokens per location
(personality & feelings)

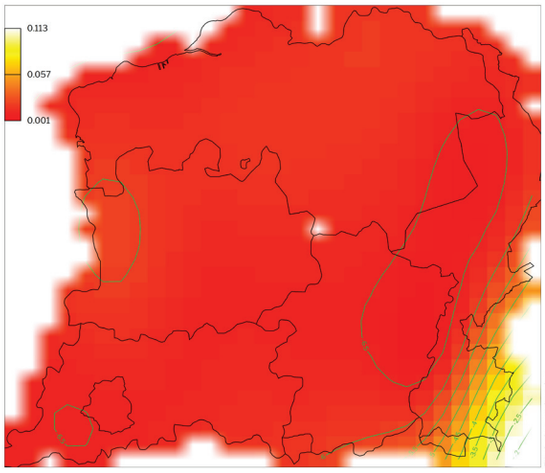


FIGURE 5.6C
German tokens per location
(church & religion)

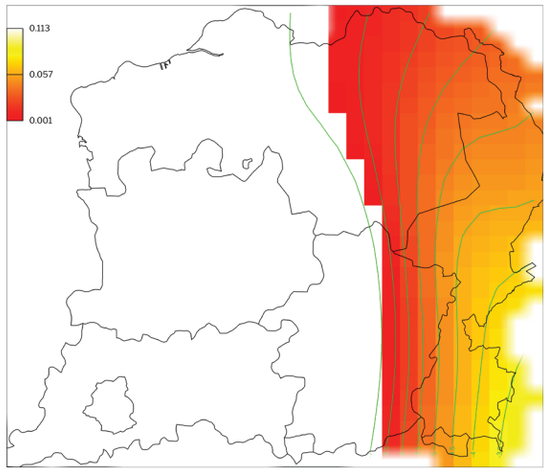


FIGURE 5.6D
German tokens per location
(clothing & personal hygiene)

dialects differ from the other Limburgish dialects due to the fact that they did undergo the second Germanic (High German) consonant shift.

Traditionally and until the beginning of the 20th century, the locations belonging to the Ripuarian dialect region in the Netherlands were oriented towards Aachen. As a result, it is not surprising that the use of German in these locations is high for the semantic fields of society and clothing, two fields that are prone to borrowing. However, Table 5.12 indicates that dialect speakers of this region also rely on loanwords related to the field of personality & feelings, which contains concepts that are generally thought to be universal and, thus, not prone to borrowing. In fact, in these locations, the proportion of German tokens for this semantic field is higher than the proportion of German for other semantic fields (it reaches an observed value of 0.18 or more).

Two interpretations for this finding can be envisaged. On the one hand, the differences between the semantic fields may reflect an older dialect situation. More specifically, since we assume that universal concepts, like those of the field of personality & feelings are not prone to borrowing (in this case from the standard variety of Dutch) and, thus, to language change at large, it may be the case that Ripuarian dialect speakers still use the old Ripuarian words, which happen to be part of the German standard language as well, as they are not marked in the dictionary as typically Ripuarian lexemes. Recall that for the French

semantic field	location (municipality)	number of German tokens	number of non-German tokens	proportion of German tokens
society, school & education	Amstenrade (Schinnen)	4	78	0.049
	Nieuwenhagen (Landgraaf)	34	595	0.054
	Mechelen (Gulpen-Wittem)	16	253	0.059
	Kerkrade (Kerkrade)	56	380	0.128
	Vaals (Vaals)	16	105	0.132
personality & feelings	Eys (Gulpen-Wittem)	93	827	0.101
	Nieuwenhagen (Landgraaf)	71	617	0.103
	Kerkrade (Kerkrade)	68	308	0.181
	Vaals (Vaals)	68	244	0.218
	Simpelveld (Simpelveld)	21	74	0.221
church & religion	Waubach (Landgraaf)	28	1005	0.027
	Chèvremont (Kerkrade)	4	91	0.042
	Bocholtz (Simpelveld)	22	384	0.054
	Montzen (Montzen)	40	580	0.065
	Kerkrade (Kerkrade)	31	415	0.070

TABLE 5.12
Locations with the highest proportion of German in three semantic fields

loanwords, we found a similar pattern (more French for personality-concepts near the border with Wallonia) and that this pattern holds when only loanwords were investigated that are not accepted in Standard Dutch. This may serve as evidence for the fact that the use of foreign material for personality-concepts in contact situations actually reflects the bilingualism of the speakers.

On the other hand, the concepts from the field of personality & feelings show a large amount of geographical variability in general. Perhaps these concepts are prime candidates for geographical variability (in this case: the use of German loanwords in a relatively limited area) because they are highly expressive. Weinreich (1968: 58-59), for instance, argues that affected concepts quickly lose their expressive meaning, which makes them prone to borrowing as language users need to be able to convey this expressive meaning. As a result, it may be the case that the dialect users rely on the German loanwords for this reason.

Tentative evidence for the second explanation comes from the fact that not every concept belonging to the field of personality & feelings for which data is available in more than one of the Ripuarian dialect locations (viz. Kerkrade, Vaals and Simpelveld), is expressed with the same German word. More specifically, for 21 out of the 95 concepts, more than one German type occurs. Furthermore, the mean proportion of German tokens per concept from this field in these locations is only 0.541, which means that only half of the "German"[8] concepts are expressed with a German token in more than one Ripuarian dialect location in the Netherlands. A BRAT (SNOTNEUS), for instance, is named a *vorwitzig* (German) in Vaals, while it is called a *muilenjan*, *snotnaas*, *kute-naas* or *kute-nelis* in Kerkrade (see Appendix

8 We use the term 'German concept' informally to indicate that the concept has at least one German token.

5.2[9]). However, for five concepts out of the 94 (viz. DECENT (*gründlich*), TO FORCE (*zwingen*), SIMPLE (*einfach*), SOBER (*einfach*) and CHASTE (*anständig*)), one German word type does seem conventionalized to some extent, as it is used in more than one Ripuarian town.

In conclusion, most of the German words for personality & feelings concepts, are not highly entrenched and conventionalized, which makes it less likely that they stem from an older language period. For only five concepts, one German word type occurs in more than one Ripuarian location. As a result, for most of the German tokens in this semantic field, the second explanation outlined above seems like the most likely one: personality & feelings concepts can be highly expressive, which results in dialect speakers relying on loanwords to convey extra (social) meaning. However, for the five concepts that do show a high amount of conventionalization, the German type may reflect an older language situation. However, additional research is necessary to corroborate these explanations further.

5.5. DISCUSSION AND CONCLUSION

The main aim of this chapter was to analyse whether semantic and lectal features affect the lexical variants that are used by a particular dialect speaker. We focused on the geographical distribution of loanwords from three different source languages and in four semantic fields. The analyses indicate that the amount of loanwords used, is highly dependent on the interaction between semantics and geography. On the one hand, we find clear differences between the source languages. Lexical borrowings from French occur the most, while German is infrequent overall. This reflects the supremacy of French culture throughout history. On the other hand, clear and systematic patterns of variation in the loanwords within one source language are significant as well. More specifically, these patterns indicate that historical and socio-cultural differences characterize the use of loanwords in the Brabantic and Limburgish dialect area.

In line with previous research, French was found to be especially frequent in the field of clothing terminology and, to a lesser extent, also in the field of society, school & education. Interestingly, the geographical pattern is not the same in both semantic fields. (Figure 5.3): the map for clothing & personal hygiene clearly shows that French is much more frequent for these concepts in Belgium than in the Netherlands. In the field of society, school & education, French additionally also occurs relatively frequently in the province of Limburg in the Netherlands. However, if the French types that also occur in the standard variety of Dutch are excluded, the distinction between the different parts of the dialect areas becomes less pronounced (Figure 5.4). The border between Belgium and the Netherlands does remain apparent.

Latin loanwords are, as expected, especially dominant in the field of church & religion. This is not surprising as many of these words were introduced as necessary loanwords: as novel concepts enter the language, the original (non-native) name for the concept is borrowed as well. However, we did find unexpected geographical differences in the spread of these Latin variants, which was tentatively explained as a result of a cultural difference between the Limburgish and Brabantic dialect area. More specifically, it may be the case that people who lived in the provinces of Limburg at the beginning of the 20th century, were more traditionally Catholic than people from the Brabantic dialect area, which is reflected in their more systematic use of the Latin variant.

German loanwords occur throughout the dialect areas in the semantic field of church & religion. Additionally, two interesting geographical patterns showed up in this and other semantic fields as well. First, for concepts from the field of society, school & education, German is only used in the Netherlands. Interestingly, we found the opposite pattern for the loanwords of French, which were more frequent in Belgium. This may indicate that French and German are complementarily distributed in this semantic field, which can be related to the fact that French culture held a stronger position in Belgium than in the Netherlands. Second, in three out of the four GAMM maps for German, the Ripuarian dialect area was clearly delineated. In this region, German tokens are always more frequent, even in the semantic field of personality & feelings. A small-scale analysis of the systematicity in the use of these German tokens in this region for personality-related concepts, revealed that the use of German is not highly systematic. Only for five concepts, a single German word type is used in every Ripuarian location. As most of the German words in the Ripuarian region are, therefore, not highly conventionalized and as many of these personality-related concepts are relatively expressive, it is possible that people living close to the German border use these words to convey extra (social) meaning.

On the basis of these source language-specific patterns, we can distinguish general implications for the borrowability of lexical material. First, in every semantic field and for every source language, systematic patterns show up that correlate to a large extent with historical evolution and the socio-cultural environment of a dialect user. As has

9 Appendix 5.2 contains an overview of the German concepts and the number of German and non-German types and tokens with which they occur in the Ripuarian dialect locations from the semantic field of personality & feelings.

been noted frequently in previous research (e.g. Backus 2013), loanword usage reflects cultural patterns. This is apparent from the different types of geographical patterns that show up. Most of these patterns can only be explained by taking into account the socio-cultural history of the dialect speakers. The influence of an exoglossic variety shows up in differences between the two countries where Dutch is spoken concerning the use of French and between the Limburgish and Brabantic language area in the use of Latin. Additionally, lexical items are also borrowed directly into the base dialects, often due to a small geographical distance to the speakers of the source language. More specifically, the border effects that were distinguished for German in the Ripuarian dialect area and for French at the Germanic-Romance language border and in the Brussels region, indicate that the distribution of loanwords in a dialect area is not only dependent on culture, but also reflects geographical closeness. In sum, the full system of loanword usage only becomes clear when the complex interaction between culture and geography is taken into account. This is in line with recent work on lexical borrowing which advocates a multifactorial approach to lexical borrowing (Zenner & Kristiansen 2014: 8-10).

Another implication for the borrowability of linguistic data is apparent from the comparison of the distribution of lexical items in the field of personality & feelings with the other semantic fields. We provide further evidence for the fact that universal concepts are less prone to borrowing. These concepts are hardly ever expressed with non-native lexical items, except in regions that have a higher degree of bilingualism, like the Ripuarian dialect area.

Finally, some scholars have argued that loanwords are copied easier from a less closely related variety (e.g. Weinreich 1968: 1-2), while lexical borrowings from a genetically close language are phonologically adapted to the language system more easily (Van der Sijs 2005: 257). We can only answer this question tentatively, because we rely on the loanword tags that are available in the dictionary, rather than interpreting the etymology of the loanwords ourselves, and because both the French and Latin culture were more important for the Belgian and Netherlandic people. The data indicate that the proportion of loanwords from Latin and, especially, from French is much higher than the loanwords from German. As a result, the data corroborate the observation that the amount of lexical items that are borrowed from a closely related language in their original form is smaller.

However, a shortcoming of this study is that the degree to which this correlation between the geographical distribution of loanwords and geographical diversity in the lexicon-at-large holds, was not taken into account directly. Complementing this research with an alternative onomasiological approach that measures variation in the success of loanword usage per concept would offer more insight into this question (see Zenner, Speelman & Geeraerts 2012). A logical extension on this study would, for instance, be to use the geographical patterns that were distinguished in this chapter and to group individual locations into large consistent, dialectal subregions. This way, the degree to which the loanwords are successful in each of these geographical subregions vis-à-vis other synonyms for a particular concept could be investigated directly. Furthermore, such a design would allow for additional, concept-based predictors to be included in the analysis as well. This could, for instance, further substantiate the interpretation of the use of German in the Ripuarian dialect area because the degree of expressivity or proneness to affect of the concepts could be taken into account as well. It would also allow for a more detailed analysis of the difference between necessary and luxury loans, which was hypothesized as an explanatory factor for the prevalence of Latin in the field of church & religion.

Another shortcoming of this study is that we did not take into account from which period a particular lexeme stems. For Latin, for instance, Van der Sijs (2005) provides a list of loanwords that were already borrowed in the Romance era, like *defensie* 'military defense' and *munt* 'coin'. These older lexical items, which have been present in the dialects for a longer time are probably more conventionalized and, thus, possibly, more widespread. As a result, it is possible that the concepts for which these types of lexemes are used, also show less lexical diversity. Furthermore, micro-level geographical patterns in the distribution of separate variants probably differ as a result of cultural or political changes as well. As a result, taking into account a factor like the age of the lexeme or concept, or comparing these relatively recent dialect data to material from an older time period, would elucidate the importance of diachronic evolution further.

Finally, we only focused on the use of borrowed material as a formal correlate of lexical diversity, but further research can be envisaged that investigates whether other naming strategies also differ as a function of historical or socio-cultural factors. Taking into account this type of variation can offer more insight into the question of how a particular lexical item becomes entrenched: if several options are available, why are some concepts expressed with loanwords in one location, while language users from a different place rely on names that are, for instance, based on a property of the referent itself, or on hyperonymic variants?

However, overall, in this chapter we were able to show that the use of loanwords varies as a function of geography and semantic field and that the patterns that we find almost exclusively reflect changes in socio-cultural history. In this chapter, we used the geographical signal in the data as the dependent variable to be explained on the basis of other

features. In the next chapter, we take a different approach. We zoom in on variation in one semantic field, the semantic field of plants, and conduct a detailed analysis of how properties of the everyday environment of a dialect speaker influence lexical variation. In this final case study, characteristics of the geographical environment of a language user will be used as an explanatory variable of lexical diversity.

6. Botany meets lexicology: The relationship between experiential salience and lexical diversity

6.1 INTRODUCTION

The first part of this dissertation indicated that lexical geographical dialect variation is influenced by cognitive characteristics of the concepts under scrutiny, related to the organization of the lexicon. In the previous chapter, we investigated how this type of variation is reflected in the use of loanwords. The analysis showed that the use of either borrowed or native lexical material for a specific concept in dialect data is dependent on an interaction between a lectal factor, viz. the geographical location of the speaker, and the meaning of the concept to be expressed. In this chapter, we further inquire into the way characteristics of the environment, including the geographical location, of a dialect speaker influence lexical diversity. In contrast with the previous chapter, however, in which socio-cultural history was used as an interpretation of the geographical patterns that occur, in this chapter we take a different approach. We determine to what extent experience-based properties, which differ between people from different locations, directly correlate with the amount of lexical diversity for a particular concept. We achieve this goal by using referential, extra-linguistic data related to the environment of the dialect speaker, as explanatory factors in the analysis. Consequently, we explicitly inquire into the degree to which the everyday environment of language users from different places influences the amount of lexical diversity.

In this final case study, the maximalist approach on meaning of the Cognitive Linguistic movement is operationalized by examining the correlation between *experiential salience*, the frequency with which a referent occurs in the natural environment of dialect speakers from different locations, and lexical diversity. Experiential salience is related to onomasiological salience, because concepts that occur more frequently in a language user's environment (i.e. concepts that are experientially salient), will probably be more psychologically entrenched as well. The relationship between experience, meaning as categorization and language is, for instance, apparent from cognitive metaphor research, especially in the context of embodiment: cross-linguistically, patterns, that are assumed to be universal, have been distinguished that explain how physical bodily processes are central to the way language users categorize their environment. Lakoff (1987: 276-278), for instance, discusses metaphors like MORE IS UP, LESS IS DOWN and PURPOSES ARE DESTINATIONS as examples of embodied categorization. These metaphors are based on preconceptual structural correlations in experience that, in turn, correlate with bodily functioning. A second strand of research that has taken into account the interplay between the environment of a language user and the language that he uses, explicitly relies on extra-linguistic, environmental properties as explanatory factors of linguistic variation or change. Studies like this are, for instance, at the core of the framework of the *Wörter und Sachen* movement (e.g. Schuchardt 1912).

The measure of experiential salience used in this chapter, however, explicitly takes into account the fact that experiences can differ between speakers. The extra-linguistic referential frequency data are geographically stratified: they take into account how often the referents occur in the environment of dialect speakers from different locations. This allows us to examine the degree to which the fact that speakers of different places encounter particular referents more (or less) frequently, which results in a higher (or lower) degree of experiential salience, correlates with geographical variation in lexical diversity. Consequently, the aim of this chapter is to examine to what degree "varying social environments may give rise to varying experiences" (Sinha & Jensen de López 2001: 20) and how this is reflected in language use.

Importantly, however, referential frequency in the environment of a language user is of course neither the only factor that correlates with language variation, nor the only aspect of experiential salience. Instead, it interacts with the social nature of language: it is not enough for a referent to frequently be present in the environment of a language user for it to become entrenched, but (names for) the referent need to be present in discourse as well. More specifically, Geeraerts (2016) argues that the degree of entrenchment of a particular linguistic item depends on three types of frequency effects that interact: experiential frequency, i.e. the frequency with which language users encounter a referent in their natural environment; conceptual frequency, the frequency with which the referent is categorized in a particular way; and lexical frequency, the frequency with which specific lexical items occur for these forms. Referential frequency, therefore, interacts with socio-cultural and linguistic factors.

On the one hand, socio-cultural practices can influence the degree to which a concept is culturally relevant, for instance, because it takes up a central position in superstition, or because it can be used as food, for medicinal practices etc. Cultural relevance can, therefore, be considered another aspect of experiential salience. Cognitive metaphor research has stressed that, although often embodied, image schemas are dependent on culture as well (e.g. Sinha 1999, Zlatev 1997). For example, in Geeraerts & Grondelaers (1995), variation in the words for ANGER is explained in terms of the historical humoral doctrine, instead of only by relying on the physiologically-based ANGER IS HEAT metaphor (from Kövesces 1989). On the other hand, linguistic factors, like concept frequency in linguistic discourse, interact with experiential salience as well. Szmrecsanyi (2016) has shown that frequency variation between the *s*- and *of*-genitive in Late Modern English not only has to do with probabilistic grammar change, but is also affected by socio-cultural variability, more specifically, changes in textual preferences, like writing less about animate noun phrases (which prefer the *s*-genitive). However, this case study complements these findings as it is the first to directly test in a systematic way to what extent experiential salience, in the form of extra-linguistic frequency counts, correlates with lexical diversity.

In practice, we inquire into the semantic field of flora and focus on variation in the names given to plants that occur naturally in the northern part of Belgium where Brabantic, Limburgish and Flemish[1] dialects of Dutch are spoken. A variety of work on the names for plants in these dialects exists (see Brok 2003), but it generally examines a different aspect of the structure of plant name variation. More specifically, most of this research focuses on a small set of plants and provides an etymological interpretation of the names that occur for these plants in different locations (e.g. Pauwels 1933, Brok 1991, 2006). Consequently, this chapter forms the bridge between the results obtained in the first part of this dissertation and the conclusions drawn in the previous chapter. In part 1 of this dissertation, we relied on internal, linguistic data to determine the degree to which a concept is onomasiologically salient (e.g. a large number of missing responses in the questionnaire for a concept indicates that the concept is less salient), which is comparable to frequency counts based on naturalistic data in (corpus) linguistic and psycholinguistic research (e.g. Divjak & Caldwell-Harris 2015). By, in this final case study, relying on referential, rather than linguistic data, we tackle another aspect of the influence of the interaction between semantic and lectal features on variation in lexical diversity. Additionally, building on the results of chapter 5, we introduce external aspects of the socio-cultural environment of a language user as explanatory factors in the analysis.

This chapter is structured as follows. Section 6.2 outlines the hypotheses of this study. Section 6.3 elaborates on the referential plant frequency data and on the linguistic data that are used. In section 6.4, the results concerning the correlation between local and global plant frequency and lexical diversity are provided. Section 6.5 provides a discussion of these results, followed by an overview of the restrictions on the present study and some suggestions for future research. Section 6.6 ties it all together in a conclusion.

6.2 HYPOTHESES

Two complementary hypotheses have been discussed concerning the influence of experiential salience on lexical diversity. On a par with the results of part 1 of this dissertation, we can expect that concepts that occur more frequently in the everyday environment of a language user and that are, therefore, experientially more salient, show a smaller amount of lexical geographical variation. More specifically, due to repeated experiential exposure, the (local) speech community may need more congruent ways of referring to the experientially more salient concepts, resulting in socio-culturally more conventionalized names and, thus, a smaller amount of lexical geographical variation. For instance, research on lexical borrowing has shown that competition between a borrowed lexeme and synonymous variants is the lowest for novel concepts: when a new concept is introduced and becomes more salient in the everyday environment of

1 Recall that we reserve the term 'Flemish' for the area where Flemish dialects are spoken, i.e. in the provinces of East and West Flanders. To avoid confusion, we do not use it to refer to the Dutch-speaking part of Belgium as a whole.

a language user, people develop a need to refer to the concept. Often, only one word for these concepts is borrowed from a foreign language (see Geeraerts 1997:108-109, Zenner, Speelman & Geeraerts 2012 and chapter 5). While the way this type of change spreads through a speech community can take place in several ways (serially or in parallel, Geeraerts 1997: 108), the underlying assumption is that, due to a larger amount of exposure to the name for the experientially salient concepts, these more frequent concepts, and their corresponding lexemes, become psychologically more entrenched.

However, a long-standing argument in linguistic theory concerning the influence of experiential (rather than *onomasiological*) salience in particular is that more salient concepts can also show *more* lexical variability. More specifically, due to a correlation between the degree of experiential salience of a concept and the amount of detail in the way the item is conceptualized, variation in the number of names for the concept can occur as well between different languages or dialectal varieties. This notorious idea was first articulated by Boas (1911), who writes that in the Eskimo languages, a larger number of names for SNOW exist than in English.[2] More specifically, categorization depends "upon the chief interests of a people; and where it is necessary to distinguish a certain phenomenon in many aspects, which in the life of the people play each an entirely independent role, many independent words may develop, while in other cases modifications of a single term may suffice" (ibid.: 26). Boas argues that this is reflected in the fact that the Eskimo people have a larger number of lexicalized words for SNOW, as they distinguish between SNOW ON THE GROUND (*aput*), FALLING SNOW (*qana*) and DRIFTING SNOW (*sirpoq*). Speakers of English, however, use only one word, *snow* (and, if necessary, linguistic strategies like modification or compounding), to characterize the (for the Eskimo people) different phenomena.

Such cultural differences can also play a role within related varieties or dialects. More specifically, according to Goossens (1964), one reason for the survival of two different names for the two handles of a scythe in the dialects spoken in the central part of the province of Limburg in Belgium, is the high frequency of usage of the instrument in this region. As a result, language users categorize the different parts of the instrument in a more detailed way, by discerning the upper from the lower handle. In the rest of the south-eastern part of the Dutch language area, his dialect map only shows one name to refer to the superordinate concept HANDLE OF A SCYTHE, which does not distinguish between the upper and lower handle. In this region, the less salient concept shows less lexical diversity.

Crucially, the difference between the two hypothesis outlined above, is situated at the "variety of variation" that they apply to (Geeraerts, Grondelaers & Bakema 1994: 1; see chapter 1). The latter hypothesis only applies to **conceptual onomasiological variation**, the situation where a particular referent can be conceptualized by means of conceptually distinct categories. For instance, to take the SNOW example first mentioned in Boas, in standard English, a smaller amount of conceptual onomasiological variation is found than in the Eskimo languages. Speakers of English all conceptualize the three types of snow, SNOW ON THE GROUND, FALLING SNOW and DRIFTING SNOW, as belonging to one category, viz. SNOW. In the Eskimo languages, however, more conceptual onomasiological variation is found: the three varieties of snow are interpreted as conceptually distinct. Crucially, these differences in the amount of conceptual variation are also reflected in the number of names that exist. In the Eskimo language, each separate snow concept comes with its own conventionalized name (viz. *aput*, *qana* and *sirpoq*), whereas in standard English, only one name occurs that applies to all the varieties of snow. The scythe example described by Goossens shows the same type of conceptual variation. For the dialect speakers of the centre of Limburg, the different types of handles of the scythe are considered to be conceptually distinct categories. For these language users, two concepts, UPPER HANDLE OF A SCYTHE and LOWER HANDLE OF A SCYTHE, exist. As a result, for each distinct concept, a separate lexical item, which is conventionalized in the local community, is available. For the speakers of the rest of the southern Dutch language area, however, the two handles are conceptualized as the same category, viz. HANDLE OF A SCYTHE. This results in only one name being available.

The former hypothesis mentioned above, that a negative correlation can be expected between experiential salience and the amount of lexical diversity, however, applies to **formal onomasiological variation**. The question then is: given a particular concept/category, how many lexical variants occur? Due to the fact that the linguistic data that we use are organized at the level of the concept, categorization differences are diminished. This is, for instance, apparent from the fact that, as mentioned in chapter 2, if the questionnaires used to elicit the data contained questions that turned out to be inadequate (for example, if the responses of the informants indicated that the questions were too specific),

2 This assertion was later taken up by Whorf (1964 [1940]: 213) as evidence for the linguistic relativity hypothesis, which entails that human beings "dissect nature along lines laid down by [their] native languages." Although this hypothesis has been open to a lot of criticism, recent studies have argued that the need for efficient communication or "cultural relativity", i.e. non-equivalence between varieties due to cultural differences (rather than on the basis of differences between language systems), may be highly relevant for cross-cultural differences in categorization, as *reflected* in language (Bromhead 2011, Regier, Carstensen & Kemp 2016, Sinha & Jensen de López 2001).

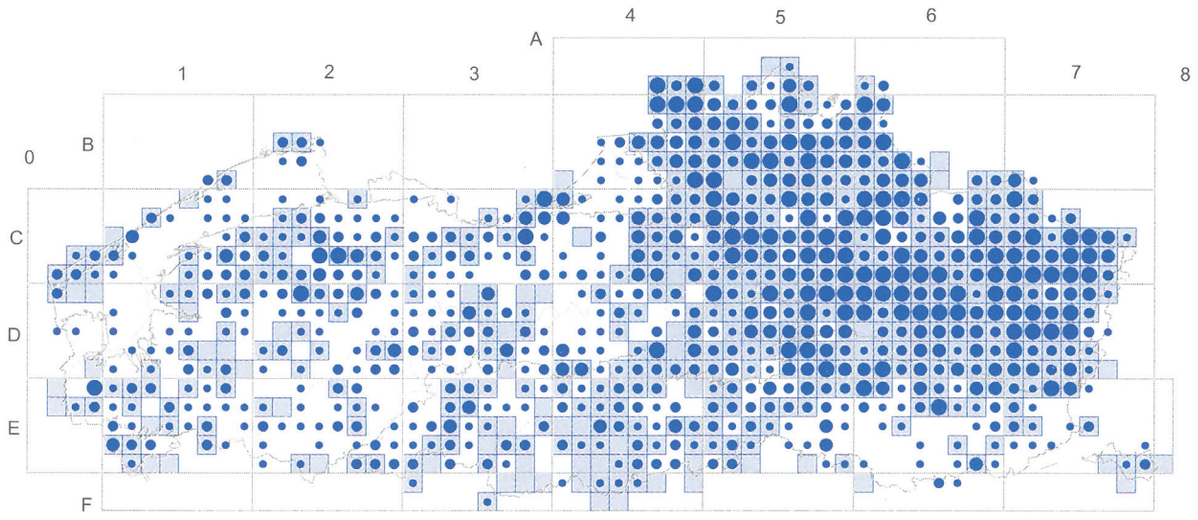


FIGURE 6.1
Geographical distribution of the common aspen (Populus tremula), a very frequent plant (Van Landuyt et al. 2006: 688).

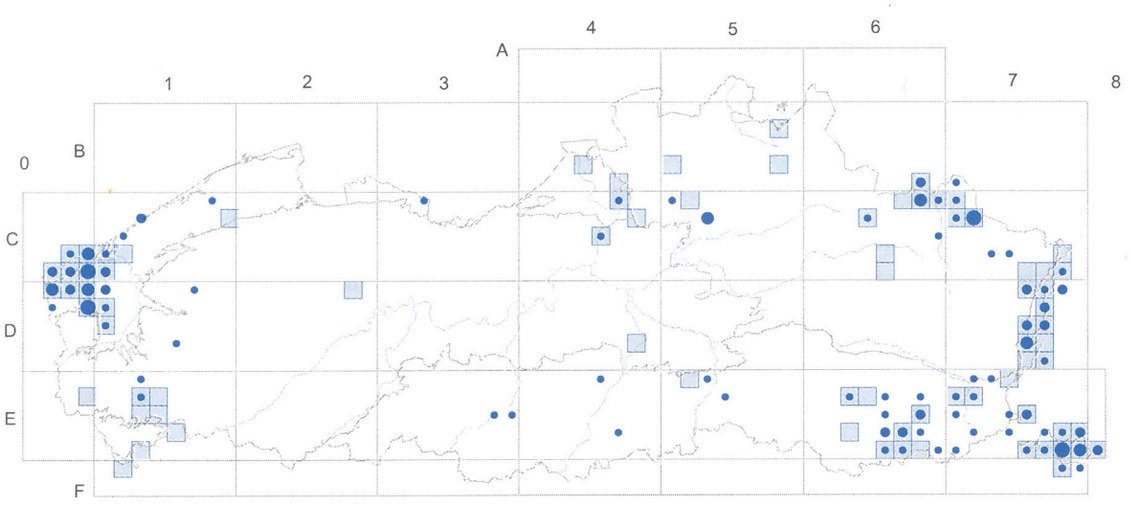


FIGURE 6.2
Geographical distribution of the common cowslip (Primula veris), a very infrequent plant (Van Landuyt et al. 2006: 712)

the questions were brought together under the same concept. For example, the overly specific questions 'shorts for boys' and 'shorts (in general)' are subsumed under the concept SHORTS. Consequently, as in this dissertation, we only inquire into variation at the level of specific concepts, i.e. formal onomasiological variation, the only hypothesis that holds is the former one: we expect to find a negative correlation between experiential salience and lexical diversity.

To test this hypothesis, we operationalize local plant salience as the frequency of the plant in the geographical area of the language user, under the assumption that plants that naturally occur more frequently in a specific region are more experientially salient for the people living in that region, because they come into contact with these plants more often. For example, salient plants, like the common aspen (Populus tremula), which grows frequently throughout the language area under scrutiny, has fewer dialectal variants in the dictionaries that we use (viz. 40) than less frequent plants like the common cowslip (Primula veris), which occurs with 217 different names. The geographical distribution of these plants is shown in Figures 6.1 and 6.2. The magnitude of the dark blue dots is proportionate to the frequency of the plant in that location (i.e. in that so-called 'hour square', see below) in the period 1972-2004. The pale blue squares reflect the distribution of the plant for the period 1939-1971 (also in hour squares). We assume that

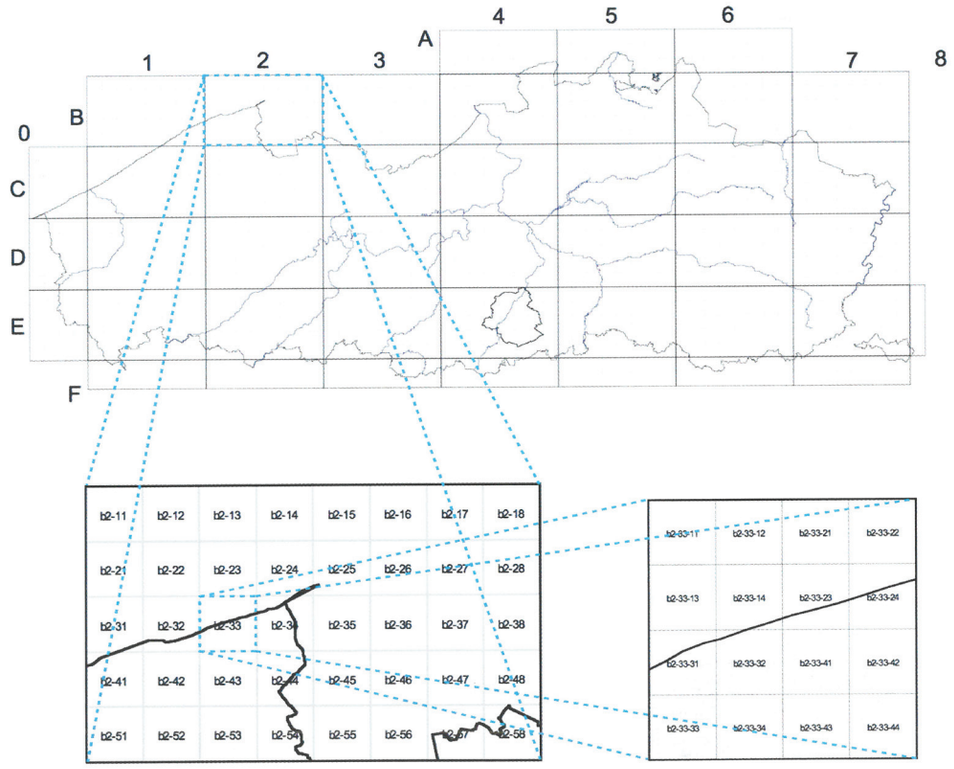


FIGURE 6.3
Hour and kilometer squares in the northern part of Belgium (Van Landuyt et al. 2006: 34)

language users refer more frequently to concepts they often come into contact with. As a result, experiential salience may affect the entrenchment and conventionalization of the names that are given to plants: experientially more salient plants are expected to show less lexical diversity.

Additionally, the fact that some plants that are infrequent in a particular region but relatively frequent across the entire language area (i.e. locally infrequent, but globally frequent) are probably better known than plants that are infrequent everywhere, may result in a higher degree of experiential salience for the first group of plants. More specifically, even though the dialect speaker does not encounter the locally infrequent plant in his everyday environment as often, he may become familiar with it when he visits peers who live farther away (where the plant is frequent), or via the media, literature or other sources. For this reason, we also take into account the global frequency of the plant in the northern part of Belgium.

6.3 DATA

6.3.1 Referential data

We use frequency data of naturally occurring plants to gauge the degree of experiential salience of the plant in the language area under investigation. These referential data come from the *Atlas van de flora van Vlaanderen en het Brussels Gewest* (Van Landuyt et al. 2006), the standard reference work concerning the distribution of plants in the northern part of Belgium. The data are also available online (http://flora.inbo.be/, Accessed on 1 August 2017).

The frequency of plants in the atlas is calculated as follows. The focus area of the atlas (i.e. the northern part of Belgium) is divided into kilometer squares of 1x1 kilometer. These kilometer squares are grouped into hour squares of 4x4 kilometers (see Figure 6.3). For each hour square, trained field workers investigated at least one quarter of the kilometer squares. The field workers were asked to record which plants they encountered while walking through the kilometer square.[3]

We adopt two types of measures of plant frequency that are available in the atlas. On the one hand, we take into account the global frequency of a plant in the northern part of Belgium, expressed as the absolute number of hour and kilometer squares where the plant was encountered. On the other hand, we also use the relative number of investigated kilometer squares in which the plant was found per ecological region to gauge the local salience of a plant. The division of the northern part of Belgium into ecological regions is based on a simplified version of the ecologically coherent

3 Some of the data in the atlas also come from secondary sources. However, for the most part, the frequency data relies on the information provided by the field workers (Van Landuyt et al. 2006:34-37).

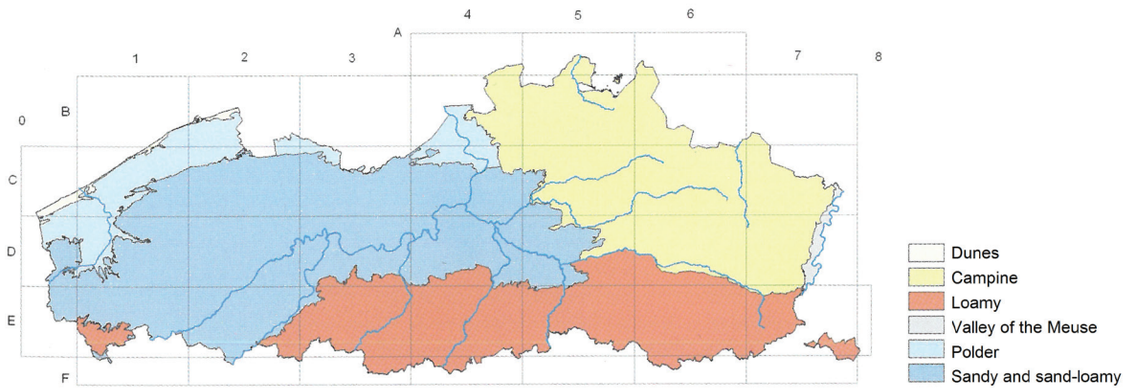


FIGURE 6.4
Ecological regions in the northern part of Belgium (Van Landuyt et al. 2006: 87)

districts described in Sevenant et al. (2002). In the atlas, six ecological regions are distinguished: the Dunes region, the Campine region, the Loamy region, the region of the Valley of the river Meuse, the Polder region and the Sandy and sand-loamy region (see Figure 6.4).

Because the atlas not only contains different measures of plant frequency (viz. local and global plant frequency), but also data from different periods[4], we use four measures of plant frequency in total: one measure of *local* plant frequency and three measures of *global* plant frequency. The measure of local plant frequency is provided as a proportion in the atlas, i.e. the number of kilometer squares in which a plant was encountered, divided by the total number of reliably investigated kilometer squares in a particular ecological region (Van Landuyt et al. 2006: 99). The measures of global frequency, however, are supplied as absolute values, i.e. the total number of kilometer or hour squares in which a plant was found in the northern part of Belgium as a whole.

1. local plant frequency: the relative number of investigated kilometer squares in which a plant was encountered per ecological region between 1972 and 2004 ('local relative frequency km squares 1972-2004')
2. global plant frequency:
    a. the absolute number of kilometer squares in which the plant was encountered throughout the northern part of Belgium between 1972 and 2004 ('global absolute frequency km squares 1972-2004')
    b. the absolute number of hour squares in which the plant was encountered throughout the northern part of Belgium between 1939 and 1971 ('global absolute frequency hour squares 1939-1971')
    c. the absolute number of hour squares in which the plant was encountered throughout the northern part of Belgium between 1972 and 2004 ('global absolute frequency hour squares 1972-2004')

As the amount of kilometer squares in the northern part of Belgium is very large and as not all kilometer squares were investigated by the fieldworkers, most plants seem to be relatively infrequent when kilometer square calculations are used (although some plants are locally very frequent, see Van Landuyt et al.: 69-80). As a result, global frequency per hour square is probably a better measure of plant frequency. However, all four plant frequency measures are highly correlated in the data set ($.85 \leq$ Spearman's rho $\leq .98$; $p < 0.001$; $N = 614$).

6.3.2 Linguistic data

The linguistic data used in this study come from three related sources. Like in the previous chapters, we use the digitized databases of the Dictionaries of the Brabantic and Limburgish Dialects (WBD & WLD). In this chapter, we focus on the *Flora* volume (III, 4.3). Additionally, we also include the database of the Flora volume of the Dictionary of the Flemish Dialects (WVD), but only the data that were collected via questionnaires. While the database of the latter dictionary is comparable to the datasets that have been used so far, some differences may occur, due to small distinctions in the history of the dictionary construction projects (Kruijsen & Van Keymeulen 1997). First, while the Brabantic and Limburgish databases are based on the same

4 Due to historical developments, the atlas contains data from two different periods (1939-1971 and 1972-2004; see Van Landuyt et al. 2006: 9-31, 35). As the data collection process has remained the same since 1939 and as we have no obvious theoretical reasons to only rely on data from one period, we include data from both periods in the analysis. Data at the level of the kilometer squares are not available for the first period (1939-1971).

ecological region	number of concepts	number of records
Dunes region	84	1887
Polder region	101	9636
Sandy and sand-loamy region	114	22755
Loamy region	132	5738
Campine region	118	692
Valley of the river Meuse	65	99

TABLE 6.1
Number of concepts and number of records per ecological region

questionnaires that were distributed by the Centre for dialectology and onomastics in Nijmegen, the questionnaires used to elicit the Flemish data, have been designed and disseminated separately by the editors of the Flemish dictionary. Second, the Flemish data were collected later (between 1998 and 2000) than the Brabantic and Limburgish data (between 1960 and 1982) and the Flemish Flora database is much larger (see below). However, as the three dictionaries have been collaborating since 1990 to achieve consistency and alignment of the databases, we believe that restricting our attention to the data that were collected through the large-scale questionnaires and to plants that occur in all three dictionaries, ensures maximal comparability between the sources. Furthermore, even though these Flemish questionnaires are not identical to the Brabantic and Limburgish ones, they are equivalent. The Flemish questionnaires include, for instance, questionnaires on plants in general (number 104, distributed in 1998), on grass (number 112, distributed in 1999) and on trees and shrubbery (number 115, distributed in 1999). The Limburgish and Brabantic data mostly come from questionnaire N 82 (1981; plants in general and trees and shrubbery), and from questionnaire N 92 (1982; names for plants and herbs).

In contrast with the previous chapters, we restrict our attention to plants that occur in all three databases and directly interlink the data from the three dictionaries semi-automatically[5] on the basis of the names for the concepts as they are presented in the dictionaries. However, as a result, it is not feasible to exclude concepts that occur in less than 50 locations and locations which have data for less than 50 concepts, because this results in too much data sparsity. Consequently, this chapter will offer further insight into the validity of the lower bound that was used in the previous chapters and into the problems and benefits of interlinking the different dialect dictionaries. The analysis will show that the amount of data is relatively small for a number of plants and differs between plants (see chapter 2 for possible explanations). We also exclude plant concepts that do not refer to actual plants, like BLOEMKNOP 'bud', or that are too general, in the sense that they do not refer to a particular type of plant, like MOS 'moss'. Overall, the linguistic dataset contains plant names collected in N = 1033 locations in the three dialect areas.

We believe that restricting our attention to the data that were collected through the large-scale questionnaires and that occur in all three dictionaries, offers two benefits. First, it ensures maximal comparability between the sources, as we investigate the same set of plants in the entire northern part of Belgium. Second, by relying on this strategy, we also have enough data at our disposal to conduct a large-scale analysis.

The final data set contains 137 different concepts. The number of concepts and the total number of records per ecological region is shown in Table 6.1. This table reveals large differences between ecological regions. On the one hand, this can be explained by the fact that the surface area of the ecological regions differs. The Dunes region, for example, is a rather narrow strip of land in the west of the Dutch-speaking part of Belgium. As a result, the number of locations in this region is relatively small. On the other hand, differences in the number of concepts and records per ecological region can also be explained by the fact that, overall, a large proportion of the data come from the WVD. This dictionary contains 30 666 records for the plant concepts under scrutiny, while the WLD and WBD combined only contain 10 203 records. As the data from the WVD mostly span the Dunes region, the Polder region and parts of the Loamy region and of the Sandy and sand-loamy region (see Figure 6.5), it is not surprising that the number of records is the largest in these regions.

Figure 6.5 further shows that the data is relatively sparse in the south of the centre of the northern part of Belgium (viz. in the province of Flemish Brabant), which is covered by the WBD. It also indicates that some locations belong to more than one ecological region. This has to do with the fact that the ecological regions are defined at the level of the municipality in Sevenant et al. (2002), even though the borders of ecological regions sometimes

5 First, we automatically match the plants that have identical concept names in the WBD and WLD with a string matching algorithm. In a second step, we manually compare the remaining plants in these dictionaries to also interlink the data that are available in both dictionaries under a different name. Finally, we add the data from the WVD to this dataset, using the same procedure.

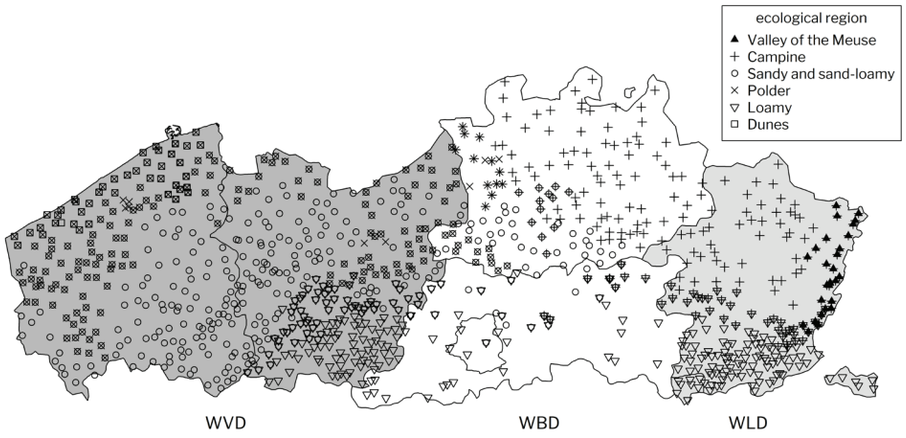


FIGURE 6.5
Dialect boundaries as represented by the WBD (white), WLD (light grey) and WVD (dark grey) and ecological regions in the northern part of Belgium

run through a municipality. For example, the municipality of Bruges belongs to three different ecological regions: the western part of Bruges belongs to the Dunes region; the central, largest part of this municipality is part of the Polder region; the eastern part of Bruges is included in the Sandy and sand-loamy region.

6.3.3 Calculating lexical diversity per concept

To operationalize the amount of lexical diversity that is found for the plant concepts in the dataset, we compare the influence of plant frequency on three measures of lexical diversity. Each measure is calculated on data containing information per plant per ecological region. The total number of concepts in the dataset is 614.

The first measure, 'number of unique types', is computed by counting the number of unique lexical items that occur per plant per ecological region. The number of unique types ranges from 1 to 92, but most concepts have a relatively low value for this variable (mean = 8.14, sd = 11.49). However, a strong positive correlation between the number of unique types and the number of records that occur in the data set per concept exists (see Figure 6.6; Spearman's rho = 0.91, $p < 0.001$). As discussed above, we only use data from the questionnaires in the dictionaries to ensure that the data was collected as systematically as possible. As Figure 6.6 indicates, however, the number of records per concept differs strongly: the number of records ranges from 1 to 4487, with mean 66.46 and standard deviation 240. Most of the concepts with a large number of records come from the Sandy and sand-loamy region (indicated with *). We expect to find negative correlations between this variable and the measures of plant frequency.

The concepts on the bottom right side of the Figure 6.6 represent the concept OAK in three different regions. From left to right, they are based on data from the Loamy region, from the Polder region and from the Sandy and sand-loamy region. This concept probably takes up a special position in this figure because of its high degree of onomasiological salience: an oak is very prototypical type of plant. Probably as a result, a large number of records for the concept are available. This high degree of onomasiological salience also explains why only a small number of unique lexical variants occur in the dataset.

A second measure of lexical diversity that is included in the analysis, is the type-token ratio (TTR) per plant per ecological region (see for example Tweedie & Baayen 1998). We use it to account for differences in the number of records (i.e. the number of tokens) that are available per concept, which can affect the number of unique types that are found for each concept per region. The type-token ratio approaches 0 when a small number of types is available, given the number of tokens. It is equal to 1 when the number of types is equal to the number of tokens.

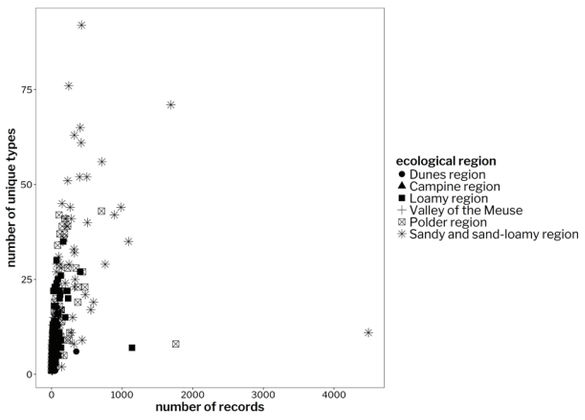


FIGURE 6.6
Correlation between number of records and number of unique types

TTR decreases when more tokens for the same number of types occur per concept, with values close to 1 expressing a large amount of lexical variation and figures close to 0 indicating that the concept shows a small amount of lexical diversity (Figure 6.7a). For example, the ratio is close to 0 when for a total of 1000 tokens, only 90 different lexical items are found (.09), while it is close to 1 when the same number of unique lexical items occurs for 100 tokens (.9). TTR is also smaller when fewer types for the same number of tokens occur per concept (Figure 6.7b), again with low values for a small amount of lexical diversity and with values close to 1 demonstrating a large amount of lexical variation for the concept. For instance, TTR is high (.9) when 90 unique lexical items occur for a total of 100 observations (i.e. a lot of lexical diversity: almost one new lexeme for every additional observation), while it is low (.1) when 10 unique lexical items occur for the same amount of tokens (i.e. little lexical variation). Consequently, we expect to find negative correlations between TTR and the plant frequency measures.

However, TTR is also sensitive to the amount of observations per concept (Speamrn's rho = -0.87, $p < 0.001$), probably because the dataset contains a relatively large proportion of concepts that have the same number of types and tokens (viz. 28.2%). For all these concepts, a limited number of records is available in the data. For example, the aspen (Populus tremula) occurs only once in the data from the Campine region; the forget-me-not (Myosotis arvensis) occurs once in the data from the Meuse valley. The plant with the largest number of types and tokens and TTR = 1 is the common corn-cockle (Agrostemma githago) in the Loamy region (11 types, 11 tokens).

A third measure we use is the measure of internal uniformity, which was first used in Geeraerts, Grondelaers & Speelman (1999; also see Speelman, Grondelaers & Geeraerts 2003) to determine the degree of uniformity in the usage of lexical variants in a speech community. We calculate this measure to determine whether plants that are more frequent in a particular region also show a higher degree of lexical standardization, in the sense that one lexical variant takes precedence over its competing heteronyms. Maximal uniformity (or standardization) occurs when everyone uses a single variant to describe a particular concept in the same situation. In our dataset, a complete lack of uniformity in an ecological region would occur when a different lexical item is used for every observation for the plant in that ecological region. However, since the ecological regions often span more than one dialect area, other factors, like dialect boundaries, probably influence the degree of standardization as well. The measure of internal uniformity ranges from 0 to 1, with 0 indicating a complete lack of uniformity (i.e. a lot of lexical diversity) and 1 indicating complete uniformity (i.e. a lack of lexical diversity). The correlation between this operationalization and the number of records per concept is lower, but still significant (Spearman's rho = -0.665, $p < 0.001$): concepts with more records in the dataset show a smaller amount of uniformity. For this variable, we expect to find positive correlations with the plant frequency measures.

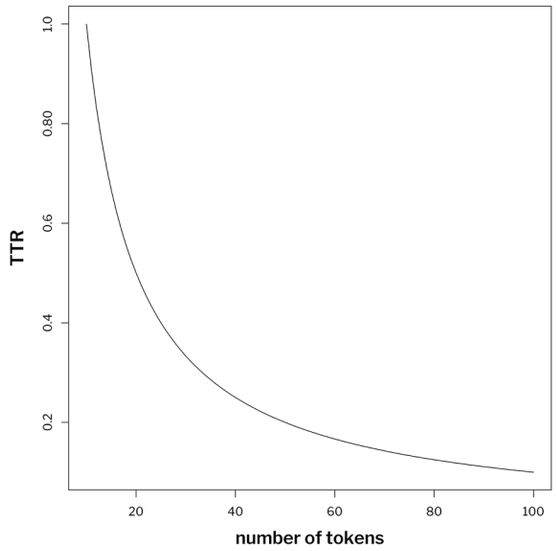


FIGURE 6.7A
Type-token ratio for increasing numbers of tokens

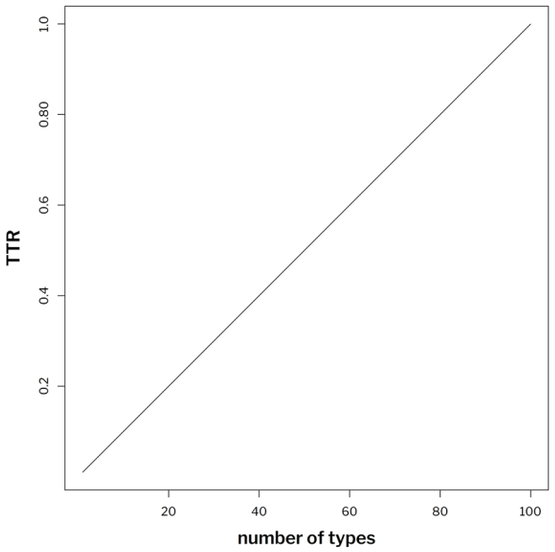


FIGURE 6.7B
Type-token ratio for increasing numbers of types

1	2	3	4	5	6	7	8	9	10	11
plant	scientific name	ecological region	nr. of records	number of unique types	TTR	internal uniformity	local rel. freq. kmsq. '72-'04	global abs. freq. kmsq. '72-'04	global abs. freq. hoursq. '39-'71	global abs. freq. hoursq. '72-'04
wood anemone	Anemone nemorosa	Campine	1	1	1	1	9.80	2031	409	507
wood anemone	Anemone nemorosa	Dunes	12	5	0.417	0.222	0.00	2031	409	507
wood anemone	Anemone nemorosa	Loamy	90	25	0.278	0.117	48.60	2031	409	507
wood anemone	Anemone nemorosa	Polder	72	23	0.319	0.118	0.40	2031	409	507
wood anemone	Anemone nemorosa	Sandy & sand-loamy	206	41	0.199	0.199	23.90	2031	409	507

TABLE 6.2
Wood anemone (Anemone nemorosa) in the final dataset; no data for the Valley of the river Meuse

To match the linguistic and the referential data, we assign each location in the dictionary data to the ecological regions that were distinguished in the atlas. For this procedure, we rely on Sevenant et al. (2002), which contains an overview of the municipalities in Belgium per ecological region. However, we make some adaptations to the description of Sevenant et al. (2002) to obtain the simplified version of the ecological regions that is used in the atlas. In a next step, we add both the global plant frequency information and local plant frequency per plant per ecological region to the dataset, on the basis of the scientific names of the plants that are provided in the Flora volume of the Dictionary of the Limburgish Dialects (p. 25-30). Finally, we calculate the number of unique types, the type-token ratio and internal uniformity per plant per ecological region on the linguistic dialect data.

For example, the dataset contains the three measures of lexical diversity (columns 5-7 in Table 6.2) for the wood anemone (Anemone nemorosa) in five ecological regions (viz. the Campine, Dunes, Loamy, Polder and Sandy and sand-loamy region; column 3).[6] It also includes the local frequency of this plant in these five regions, expressed in percentages (column 8), and the global frequency of the plant (measured in three ways in columns 9-11, see above) in the northern part of Belgium. In contrast with the measure of local frequency (i.e. per ecological region), global plant frequency is the same in every ecological region, as it is a measure of the frequency of the plant in the northern part of Belgium as a whole. In the analysis, we aggregate over all the regions and over all the plants (N = 614). We test whether the measures of lexical diversity (columns 5-7) correlate with the measures of plant frequency (columns 8-11).

6.4 ANALYSIS AND RESULTS

This section presents the analyses and the results. In 6.4.1, we correlate the four measures of plant frequency from the atlas with the three measures of lexical diversity, calculated per plant per ecological region. In 6.4.2, the relationship between global and local plant frequency is scrutinized. Importantly, the interpretation of the results differs for the number of unique types per concept and TTR on the one hand, and for internal uniformity on the other hand. A positive correlation for plant frequency and the former measures indicates that more frequent plants show more lexical diversity. However, a positive correlation coefficient for plant frequency and the latter measure shows that internal uniformity correlates positively with plant frequency and, thus, that more frequent plants show less lexical diversity. An explanation for the findings outlined below is provided in the discussion (6.5).

6 We have no information about the wood anemone in the ecological region of the Valley of the Meuse, because no linguistic data is available for this plant from locations belonging to this region.

	number of unique types	type-token ratio (TTR)	internal uniformity
local relative frequency km squares 1972-2004	0.261 p < 0.001	-0.256 p < 0.001	-0.191 p < 0.001
global absolute frequency km squares 1972-2004	0.241 p < 0.001	-0.261 p < 0.001	-0.156 p < 0.001
global absolute frequency hour squares 1939-1971	0.233 p < 0.001	-0.223 p < 0.001	-0.155 p < 0.001
global absolute frequency hour squares 1972-2004	0.240 p < 0.001	-0.256 p < 0.001	-0.158 p < 0.001

TABLE 6.3
Spearman's rank correlation coefficient and corresponding p-value for the relationship between measures of plant frequency and measures of lexical diversity per plant

6.4.1 The relationship between plant frequency and lexical diversity

Correlating plant frequency and lexical diversity

To test whether plant frequency has a significant influence on the diversity in the names for plants in the data set, we use Spearman's rank correlation tests. More specifically, we test whether the plant frequency measures ('local relative frequency km squares 1972-2004', 'global absolute frequency km squares 1972-2004', 'global absolute frequency hour squares 1939-1971' and 'global absolute frequency hour squares 1972-2004') correlate significantly with each of the three operationalizations of lexical diversity ('number of unique types', 'type-token ratio' and 'internal uniformity'). We also calculate the correlation coefficient using Spearman's rank correlation tests. This coefficient ranges from -1 to 1, with negative values representing a negative correlation between the variables and positive values indicating a positive correlation. When the coefficient is 0, no correlation between the variables is found.

	correlation coefficient and p-value for Spearman's rank correlation
local relative frequency km squares 1972-2004	-0.220 p < 0.001
global absolute frequency km squares 1972-2004	-0.261 p < 0.001
global absolute frequency hour squares 1939-1971	-0.206 p < 0.001
global absolute frequency hour squares 1972-2004	-0.255 p < 0.001

TABLE 6.4
Correlation between four measures of plant frequency and TTR per plant for concepts with TTR smaller than 1 (N = 441)

Results

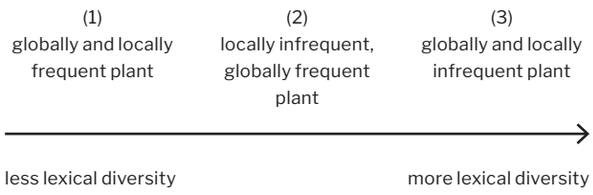
Table 6.3 provides an overview of the p-value and Spearman's rank correlation coefficient for each combination of the measures of plant frequency and of lexical diversity per plant (N = 614). The table indicates that a significant correlation (alpha = 0.05) exists between plant frequency and lexical diversity in all the cells. However, as the absolute values of the coefficients are never larger than 0.261, the correlation between plant frequency and lexical diversity is not very strong. Furthermore, plant frequency does not always correlate with lexical diversity in the way that was expected. More specifically, for 'number of unique types', which is shown in the second column of the table, positive correlations are found, while internal uniformity, in the fourth column, shows significant negative correlations. This means that more variation is found for plants that are more frequent, both locally and globally.

For the third column in the table, which provides the results for TTR, all the measures of plant frequency show a negative correlation with lexical diversity. This is in accordance with what was expected: more frequent plants show a smaller amount of lexical diversity. However, as suggested in section 6.3.3, TTR is sensitive to the number of tokens per concept, in the sense that TTR is high for concepts with the same number of types and tokens, even when only a small number of records is available for these concepts. As about one third of the concepts in the data set have a TTR value of 1, inspecting whether the negative correlation persists when only plants with a TTR value lower than 1 are

included in the analysis may offer some more insight into the relation between plant frequency and TTR. Table 6.4 shows Spearman's rank correlation coefficients and the p-values for this subset of the data (N = 441). Even though the correlation coefficients are slightly lower than in Table 6.3, the significant negative correlations persist: more frequent plants show a smaller amount of lexical diversity.[7]

6.4.2 The relationship between the local and global frequency of a plant

Concerning the relationship between the four measures of plant frequency that were used, Tables 6.3 and 6.4 show that both the local and global frequency of a plant correlate with lexical diversity. By solely relying on these measures, we cannot determine whether local and global frequency have the same effect on lexical diversity. As explained at the end of section 6.2, we assume that plants that are only infrequent in a particular region are still more salient overall than globally infrequent plants and, thus, show less lexical diversity.[8] In sum, we expect to find that lexical diversity follows the following pattern:



Comparing local and global plant frequency

To determine whether this relationship holds, we build three mixed-effects linear regression models with as a response variable the number of unique types (model 1), TTR (model 2) and internal uniformity (model 3) per plant per ecological region. Since the dataset contains multiple observations for each ecological region and for most of the plants, we use these factors as random effects in the models. We include frequency category per plant as a fixed-effects predictor in each of the models (N = 336).[9] This variable has three possible levels, depending on the global and local frequency of the plant:

1. very frequent plants, i.e. plants that occur in at least 2/3 of the hour squares that were investigated between 1939 and 1971 and that are available in at least 70% of the kilometer squares of the region under scrutiny (N = 106), e.g. the common nettle (Urtica dioica) in all ecological regions;
2. plants that are globally frequent, but infrequent in a particular region, i.e. plants that occur in at least 2/3 of the hour squares that were investigated between 1939 and 1971, but that are only available in less than half of the km squares in a particular region (N = 51), e.g. the common bent (Agrostis capillaris) in the Polder region;
3. plants that are globally and locally infrequent, i.e. plants that occur in less than 1/3 of the hour squares that were investigated between 1939 and 1971 and that are only available in less than half of the km squares in a particular region (N = 179), e.g. the sweetscented bedstraw (Galium odoratum) in all ecological regions.

Results

Table 6.5 shows the output of the three regression models. At the top of the table, the random effects (all adjustments to the intercept) are shown with their corresponding standard deviation, and the residual error. Each model has the same random effects structure, with a random intercept for plant and a random intercept for ecological region. In each of the models, this random structure was statistically validated before including the fixed-effects predictor.[10] The bottom of the page shows the model diagnostics. Marginal and

7 Additionally, we checked whether significant correlations are also found for plant concepts with at least 50 records. This data set is smaller (N = 137) and, probably as a result, some of the plant frequency measures lose their significance. All the correlation coefficients have the same sign as in the larger data set, but the absolute values are lower. A significant correlation is still found between TTR and 'global absolute frequency hour squares 1972-2004' (Spearman's rho = -0.18; $p < 0.05$). Near-significant negative correlations, which would probably reach significance in a larger data set, still occur between TTR and 'global absolute frequency km squares 1972-2004' (Spearman's rho = -0.16; $p < 0.1$), and between internal uniformity and 'local relative frequency km squares 1972-2004' (Spearman's rho = -0.15; $p < 0.1$). Overall, these results suggest that the relationship between plant frequency and lexical diversity is not solely dependent on the amount of data available per concept.

8 Differences in experiential salience may also be found when a plant is locally frequent but globally infrequent. However, as only two plants in our dataset would belong to this category (viz. the wild privet (Ligustrum vulgare) and the goldmoss stonecrop (Sedum acre), two plants that are typically found near the sea and, thus, grow frequently in the Dunes area, but only rarely occur naturally in the rest of the northern part of Belgium), we do not take this category into account.

9 Because we are mostly interested in the extreme cases in this part of the analysis, we do not include all the plants in the models. More specifically, plants that are relatively 'neutral' regarding global or local frequency, i.e. plants that are neither locally, nor globally very frequent or infrequent, are not assigned to any of the frequency categories.

10 Ideally, we would have liked to use a random intercept for each plant per ecological region. However, the data do not support models with a random structure this complex. Instead, we use a separate random intercept for plant and ecological region and verify that intercept-only models with this random structure perform better than models without one or both of these random intercepts.

	model 1 nr. of unique types			model 2 TTR			model 3 internal uniformity		
random effects									
			std. dev			std. dev			std. dev
plant	intercept		6.240	intercept		0.210	intercept		0.133
ecological region	intercept		6.131	intercept		0.204	intercept		0.188
residual			8.593			0.201			0.250
fixed effects									
	estimate	SE	p-value	estimate	SE	p-value	estimate	SE	p-value
intercept (glob. freq.)	11.004	2.875	< 0.01	0.434	0.093	< 0.01	0.4701	0.0842	< 0.01
locally infrequent	-1.549	2.158	NS	0.111	0.058	<0.1	0.1183	0.0564	< 0.05
globally infrequent	-7.035	1.808	< 0.001	0.243	0.054	< 0.001	0.1617	0.0443	< 0.001
model diagnostics									
marginal R^2	0.068			0.087			0.043		
conditional R^2	0.542			0.707			0.481		

TABLE 6.5

Output for the random and fixed effects for mixed-effects linear regression models with as response variables the number of unique types per plant (model 1), TTR per plant (model 2) and internal uniformity per plant (model 3) in function of plant frequency category (reference level: globally frequent plants). Marginal R^2 shows the proportion of variance explained by the fixed effects alone. Conditional R^2 depicts the proportion of variance explained by the fixed and random factors.

conditional R^2 show the proportion of variance explained by the fixed effects alone, and the proportion explained by the combination of the fixed and random factors, respectively.[11]

The middle part of Table 6.5 shows the estimate and p-value for the fixed-effects predictor 'frequency category'. In models 1 and 2, a higher value for the response variable indicates a larger amount of lexical variation per plant (operationalized as number of unique types and TTR, respectively). In these models, we would therefore expect positive estimates for the locally and globally infrequent plants, in comparison to the reference level (globally and locally frequent plants), which is captured in the intercept. However, the results are not completely in line with this expectation. For number of unique types, the amount of variation decreases for less frequent plants. However, this unexpected negative trend is probably connected to the fact that for less frequent plants, a smaller amount of records is available per plant. In fact, there is a significant positive correlation between the number of responses per plant and the three plant frequency categories ($H = 31.645$, $p < 0.001$). The globally frequent plants have 160 records on average (sd = 497); for locally infrequent plants, the mean number of records per plant is 93 (sd = 259); for globally infrequent plants, the average number of records is only 26 (sd = 60). As the number of lexemes and the number of records per plant per region are highly correlated (see 6.3.3), it is not surprising that for the less frequent plants, a smaller number of unique types is found.

11 Marginal and conditional R^2 were calculated using sem.model.fits() from the piecewiseSEM-package (see https://jonlefcheck.net/2013/03/13/r2-for-linear-mixed-effects-models/, Accessed on 5 May 2017).

	plant name, ecological region	number of records	distribution of types	number of unique types	TTR	internal uniformity
1	great mullein (Verbascum Thapsus), Loamy region	26	$lexeme_{1...18}$ occur once $lexeme_{19...22}$ occur twice	22	0.846	0.050
2	bitter dock (Rumex obtusifolius), Polder region	38	$lexeme_{1,2}$ occur once $lexeme_{3}$ occurs 3 times $lexeme_{4}$ occurs 4 times $lexeme_{5}$ occurs 10 times $lexeme_{6}$ occurs 19 times	6	0.158	0.338
3	black locust (Robinia pseudoacacia), Sandy and sand-loamy region	26	$lexeme_{1,2,3}$ occur once $lexeme_{4}$ occurs 23 times	4	0.154	0.787
4	forget-me-not (Myosotis arvensis), Dunes region	52	$lexeme_{1}$ occurs 52 times	1	0.019	1

TABLE 6.6
Comparison of number of unique types, TTR and internal uniformity

plant	ecological region	number of records	number of unique types	TTR	internal uniformity
broadleaf plantain (Plantago major)	Sandy and sand-loamy	218	39	0.179	0.079
lesser burdock (Arctium minus)	Sandy and sand-loamy	420	61	0.145	0.100
blackberry bush (Rubus fruticosus)	Sandy and sand-loamy	500	52	0.104	0.106
English plantain (Plantago lanceolate)	Sandy and sand-loamy	141	28	0.199	0.111
lesser burdock (Arctium minus)	Polder	226	39	0.173	0.112

TABLE 6.7
Overview of the five plants with the lowest value for internal uniformity and TTR < .2

In model 3, higher values for the response variable signify a smaller amount of variability. We therefore expect negative estimates for the locally and globally infrequent plants. However, in this model, we find the opposite effect as well: less frequent plants show a significantly larger amount of internal uniformity. In sum, only the results for TTR are as expected: both locally and globally infrequent plants have a significantly higher estimate than the frequent plants. This means that the less frequent a plant is, the higher its TTR value and, thus, the larger the amount of variation in the names for the plant.

6.5 DISCUSSION

Overall, the results of our analyses show that a correlation exists between plant frequency and lexical diversity. Although we aimed to show that experientially more salient

lexical item	N	lexical item	N	lexical item	N
kleef	2	plakkerbollen	2	plakbollen	4
klitkruid	2	plakkersbezetjes	2	plakdistel	4
wier	2	plakkerstruik	2	plakkers-, plakkertjeskruid	4
bommetjes	2	plakmadammetje	2	plakmadammetjes	4
bot	2	plakt-de-baard	2	distel	6
distelknoop	2	reit	2	klit	6
distelstekker	2	smijtdodde	2	distels	6
distelvinken	2	smijters	2	plakker	6
doppers	2	speenkruid	2	klis(se)bol	8
dotsjes	2	stekelharen	2	soldate-, soldatenknop(je)	8
everzwijnkruid	2	stekeltjes	2	klis(se)kruid	10
haakbloemen	2	stekers, stekertjes	2	stekkers, stekkertjes	12
klauwkruid	2	stekker	2	plakkruid	14
kleeftebollen	2	stekkertjeskruid	2	plakkers, plakkertjes	14
klissenstok	2	sterkerbol	2	kleefte	20
klister	2	toorvel	2	klissen	26
knopkruid	2	weerhaakjes	2	soldate(n)knoppen	28
mottebollen	2	zoete distel	2	kleef-, klevekruid	34
mouwenkruipers	2	grote klis	4	klis	116
pieker	2	kleefbollen	4		
piekertjes	2	klissebollen	4		

TABLE 6.8
Frequency of lexical items for the lesser burdock in the Sandy- and sand-loamy region (N = 420).

plants show less lexical diversity, the results are not completely in line with this expectation. One explanation for this finding is that the correlation between the measures of lexical diversity and the number of records that are available per plant, influences the results to a certain extent, especially for the number of unique types (see 6.3.3). Interestingly, the results for TTR and internal uniformity also differ, even though both of these measures take the number of tokens per concept into account. Before identifying some suggestions for future research in 6.5.3, section 6.5.1 will outline two explanations for these diverging results. On the one hand, TTR and internal uniformity can be different because they measure conceptually different phenomena. On the other hand, the measures were calculated per ecological region, but an ecological region may include different dialect regions. In 6.5.2, the direct interlinking of the three dialect dictionaries is discussed.

6.5.1 TTR versus internal uniformity

The results for the TTR measure are as expected (less lexical diversity is found for more frequent plants and locally infrequent plants show less lexical variation than globally

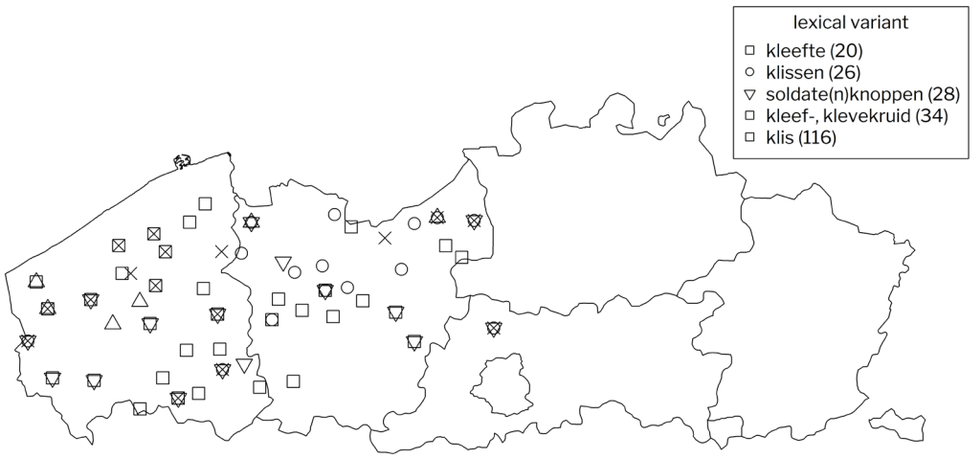


FIGURE 6.8
Geographical distribution of lexemes with N ≥ 15 for the lesser burdock in the Sandy and sand-loamy region

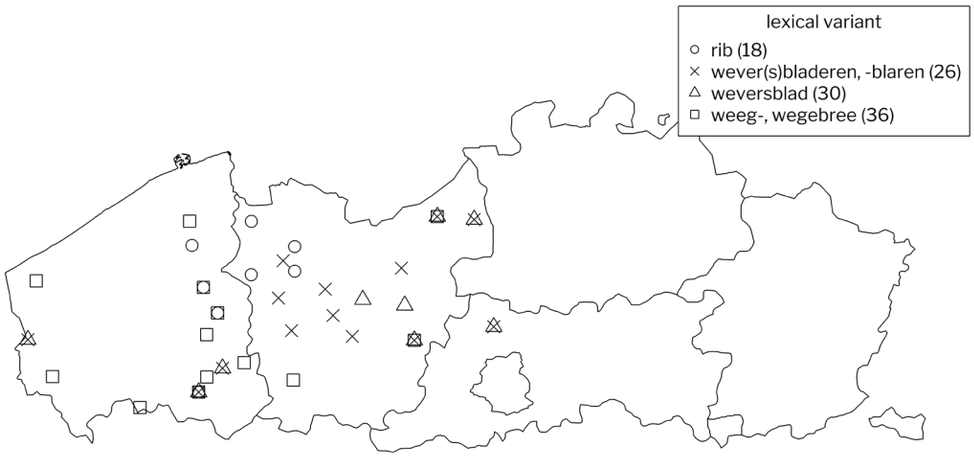


FIGURE 6.9
Distribution of lexemes with N >= 15 for the broadleaf plantain in the Sandy and sand-loamy region

infrequent plants). Furthermore, the correlation persists even when only concepts are included in the analysis for which TTR is smaller than 1 (see Table 6.4). The results for internal uniformity show the opposite trend. Because the measures of lexical diversity are calculated at the level of the ecological region, the relationship between internal uniformity and TTR can probably be explained in terms of the degree of standardization per ecological region.

Table 6.6 shows the difference between the two measures. The number of records is comparable for the four plants, great mullein (Verbascum Thapsus) in the Loamy region, bitter dock (Rumex obtusifolius) in the Polder region, black locust (Robinia pseudoacacia) in the Sandy and sand-loamy region and forget-me-not (Myosotis arvensis) in the Dunes region. The number of unique types decreases from top to bottom (see Appendix 6.1 for an overview of the lexical items used per plant). Table 6.6 confirms that while the TTR measure cannot distinguish row 2 from the third one, the measure of internal uniformity can. The latter is sensitive to the number of lexemes that occur per concept *and* to the number of tokens per lexeme (i.e. type). It is low for concepts which show a smaller amount of standardization (i.e. one lexical item takes precedence over its competing dialectal heteronyms), like the bitter dock in the Polder region, and higher for plants with a larger degree of standardization, like the black locust in the Sandy and sand-loamy region.

As a consequence, even though plant frequency has an influence on the number of lexemes per concept, as indicated by the results for TTR, it does not necessarily ensure that one lexeme becomes the preferred lexeme over its competing

lexical item	N	lexical item	N	lexical item	N
bree	2	varkensblad	2	zwijnegras	2
zwijnsoren	2	varkensblaren	2	grote weegbree	4
boterblad	2	varkensgras	2	kattestaart	4
breedblad	2	weegiebladen	2	weeg-, wege(s)bladen, -blaren	4
breedbladige weegbree	2	weegweeblad	2	wegaard(s)blad	4
breedbladweegbree	2	weewaarsblad	2	wegbree	6
dokke	2	weeweeblad	2	weewaarsbladen	8
dokkeblaren	2	weeweegbree	2	honde-, hondsrib	10
grote smart	2	wegaardsblaren	2	brede weegbree	14
honderibben, hondsribberen	2	wemel	2	rib	18
keunoren	2	weversbloemen	2	wever(s)bladeren, -blaren	26
papbladen	2	wilgebladen	2	weversblad	30
platen	2	zevenblaren	2	weeg-, wegebree	36

TABLE 6.9
Frequency of lexical items for the broadleaf plantain in the Sandy- and sand-loamy region (N = 218).

synonyms throughout the ecological region. While for more frequent plants, the number of different variants decreases for the same amount of tokens, this does not mean that every language user chooses the same name in the same situation (i.e. ecological region). Geographical variation within an ecological region, for example, is not neutralized by the high natural frequency of a plant. In fact, if a plant has both a low value for TTR and for internal uniformity, this means that, while the plant does not have a large number of unique types given the number of available tokens, the number of records per lexeme per plant per region does not differ a lot and the tokens are distributed over the unique types in a relatively homogeneous way.

By inspecting the frequency of the lexemes for globally frequent plants with both a low value for TTR and for internal uniformity, we can confirm whether this explanation holds. Table 6.7 shows the five plants with the lowest value for internal uniformity and TTR < 0.2. Tables 6.8 and 6.9 show the frequency of the lexical items that are used for the lesser burdock and the broadleaf plantain in the Sandy and sand-loamy region, (rows 1 and 2 in Table 6.7) which will be discussed in more detail below. The distribution of the lexemes for the other plants in Table 6.6 is comparable to these plants (see Appendix 6.2): all five plants have about 3-5 lexemes that are very frequent in comparison to the other words for the concept.

For the lesser burdock in the Sandy and sand-loamy region (N = 420), for example, *klis* occurs 116 times (see Table 6.8). Four other lexemes occur more than 15 times (*kleefte*; *klissen*; *soldate(n)knoppen* and *kleef-, klevekruid*). The other lexemes are less frequent. Overall, the tokens of these plants are distributed in a relatively homogeneous way over the unique types. Plotting the geographical distribution of the lexemes on a map indicates that more than one lexeme occurs in some locations: the language users know more than one local dialect word to refer to the concept (Figure 6.8). *Klis* is used throughout the ecological region. Other variants sometimes occur in locations where *klis* was found as well, or in locations close to places with *klis*. Interestingly, these other variants also have a more limited geographical distribution than *klis*.

Furthermore, other factors can be envisaged that determine which lexeme is used in which location. For example, it may be the case that the geographical distribution of the variants within the ecological regions reflects dialect boundaries and, thus, does show some degree of standardization, albeit on a different level than per ecological region.

In this case, one would be able to find a number of relatively small geographical areas where a particular variant is used. An example of this can be found if the variants for the broadleaf plantain that occur more than 15 times in the data are plotted on a map (Figure 6.9, also see Table 6.9). Even though these variants are relatively frequent in comparison to the other lexemes for this concept, they all seem to only be used in a particular geographical area of the Sandy and sand-loamy region.

The diverging results for the models for TTR and internal uniformity per plant frequency group (Section 6.4.2) can be interpreted in a similar way. The analysis showed that the predicted value for TTR and for internal uniformity is smaller for the very frequent plants than for the locally and globally infrequent plants. The smaller values for TTR are in line with what was expected: a high value for global frequency can reduce the amount of diversity in the names for locally infrequent plants. The results for internal uniformity seem to contradict this finding. However, it is possible that the unexpected lower degree of uniformity of frequent plants is again related to the fact that there is no uniformity within the ecological region: the tokens for these plants may be distributed among the unique types that occur for the plants in a relatively homogeneous way. Additionally, since the number of records per plant also correlates with the frequency of the plant, a smaller number of tokens (and, thus, types) is available for the infrequent plants. This results in a seemingly more homogeneous distribution of the variants in the ecological regions (high degree of internal uniformity) and in a higher value for TTR.

6.5.2 Combining the three regional dialect dictionaries

The previous paragraph showed that using calculations of lexical diversity on a different level than per dictionary, may obscure geographical variation within an ecological region. Furthermore, in contrast with the previous chapters, we used a different method to subset the source data from the WBD, WLD and WVD. The analyses show that this is not unproblematic.

First, we did not set a lower bound on the number of concepts per location and on the number of locations per concept that have to be available. This results in significant positive correlations between the measures of experiential frequency and the amount of records that are available per plant. For the measure of number of unique types, this is particularly problematic, because this variable is heavily influenced by the number of available records. On the one hand, this produces an unexpected sign for the correlation between experiential salience and this measure of lexical diversity in the dataset: less frequent plants show less variation, probably due to the fact that fewer records are available. Crucially, however, in contrast with the results for TTR (and internal uniformity), these correlations lose their significance if enough data is taken into account (see 6.4.1 and footnote 7). Consequently, setting a lower bound on the number of concepts and locations is necessary if the analysis relies on a measure of lexical (geographical) variation that does not take into account the number of tokens that are available per concept. Since the response variable in chapters 3 and 4 also depends on the number of unique types per concept, using the lower bound in the other case studies clearly improves the validity of these analyses. On the other hand, the correlations between the measures of experiential frequency and the amount of records per plant corroborate that, at least in the Flora issue of the dictionaries, the respondents provided fewer dialectal variants for the less experientially salient concepts in the questionnaires. This is important, because it serves as evidence for the validity of operationalizing lack of salience as the proportion of missing places per concept. Although this variable was shown to be problematic in chapter 3 and 4, the results of the current chapter indicate that less well-known plants occur with fewer responses in the database.

Second, we lumped together all the data from the three dictionaries that were used in the analyses. However, as outlined in chapter 2, this means that we are assuming a certain degree of homogeneity in the dataset that may not always be present. For instance, although the data from the three dictionaries were not collected in exactly the same period, we did not control for diachronic differences between the sources: because most of the data come from the dictionary of Flemish dialects and because we aggregate over all the plants and ecological regions, we expect that this diachronic noise does not bias the analysis to a large degree. Further, since the editors of the three dictionaries probably did not always make the same decisions about how to group different phonological variants into one lexeme, the data set may contain false heteronyms, lexemes that are treated as separate headwords in one dictionary, while they are treated as the same word in another one. For example, in the WLD, the phonological variant *bosbessen* 'bilberry' is grouped under the lexeme *bosbes*, while in the WBD, related phonological variants like *bosbeize*, *bosbeze* and *bosbieseme* are grouped under *bosbezen*, *bosbezen* and *bosbezem*, respectively. To cope with this difficulty, it would be necessary to compare the dutchified lexical items for the phonological variants in all the dictionaries. However, as this dissertation aims to take a large-scale approach towards lexical geographical variation, we assume that the dictionaries are similar enough to be compared and that this kind of noise is filtered out due to the aggregate approach that we employ. Therefore, an interesting addition to this study would be to extend the scope

to other dialect or language areas to investigate whether the findings are stable in other datasets and outside the region of the northern part of Belgium.

6.5.3 Suggestions for future research

First, as we do not find a high degree of lexical homogeneity per ecological region, other operationalizations of the division into geographical regions could be used. An alternative approach to examining the correlation between referential plant frequency and lexical diversity could, for instance, take the form of investigating geographical areas that are defined on the basis of linguistic data, rather than on the basis of the abiotic features, like humidity or type of soil, that underlie the ecological regions. For instance, one alternative approach would be to rely on descriptions of the Dutch language area into smaller dialect regions and to calculate the measures of lexical diversity for each of these regions (e.g. using the map of Daan (1969) or the divisions available in the WBD, WLD & WVD). However, some of these linguistically defined regions would probably be characterized by different types of ecological systems. The Limburgish town of Diepenbeek, for instance, is ecologically relatively diverse as its southern part belongs to the Loamy region, while the northern part belongs to the Campine region. Additionally, at this point, a practical issue prevents us from conducting such an analysis. More specifically, the referential data from the atlas are only available in an aggregate form (i.e. in the form of frequency counts for the entire region of Belgium or for an entire ecological region). Quantitative data of the presence of a particular plant in a particular kilometer or hour square cannot be as easily obtained.

Second, the analysis also showed that the absolute value of the correlation coefficients is relatively low (it is never higher than 0.261). This confirms the observation already mentioned in the introduction, that other factors than referential frequency interact with experiential salience and probably influence the amount of lexical diversity found in names for plants. More specifically, lexical variation depends on the social nature of language: it is not enough for a referent to frequently be present in the environment of a language user for it to become entrenched, but (names for) the referent need to be used in discourse as well.

An additional explanation for the low correlation coefficients in the analysis is that the plants that are included in the dictionary data are overall relatively frequent. The mean value for relative frequency per ecological region per plant for all the plants in the online database of the atlas is 12.46%. The mean value for this measure in the data set that was used for this paper is 37.78%. Of course, it is not surprising that only dialect data for relatively frequent plants is at our disposal. On the one hand, some of the plants in the atlas are probably so infrequent that they are not known to laymen. As a result, it may be the case that the lexicographers are not aware that these plants exist. On the other hand, if they are aware of the plants, it is possible that they are not interested in the names for these plants in local dialects, because they expect that asking for the names for these plants will not provide them with enough data. As was shown above, even for the relatively frequent plants that are available in our dataset, some plants are not represented by a large number of records in the linguistic data, which may have to do with the fact that these plants are unfamiliar for language users. Collecting dialect data for less frequent plants could corroborate the findings of this paper further.

Follow-up research, should, thus, compare alternative operationalizations of the response and predictor variables to the measures that were used here. For instance, using a response variable that directly takes into account the amount of geographical variability per concept (as we did in the other chapters), could further corroborate the assumption that geographical variation within an ecological region may explain some of the conflicting results between TTR and internal uniformity. Extensions of the predictor variable, experiential frequency, are possible as well. For example, the poisonousness, usefulness or folkloric salience of a plant can also influence how familiar the plant is. Additionally, in this chapter we assume that the degree to which a plant is experientially salient affects lexical geographical variation in the same way that onomasiological salience does. However, on the basis of these data, the precise relationship, and the process by which an experientially salient object becomes onomasiologically entrenched cannot be distinguished. Other types of data, like psychometric ratings or experimental results may offer further insight into this question. Furthermore, comparing lexical data across different time periods can reveal whether the degree of lexical diversity decreases for plant names over time, and whether this is influenced by plant frequency.

Finally, in this chapter we aimed to investigate whether experiential salience, in the form of referential frequency, influences lexical diversity. Other semantic fields can be envisaged in which this correlation can be tested. For example, rather than focusing on flora (or fauna), it would be interesting to expand the scope to a semantic field that is more prone to cultural differences, like the field of artefacts. Using other semantic fields will also allow for a comparison between concepts that occur naturally or that are conceived in a social environment.

6.6 Conclusion

In this paper we linked referential data to linguistic data to test whether the referential frequency of a plant, which was used to gauge experiential salience, correlates with the amount of lexical variation that is found in the names for the plant. The analysis showed that some significant correlations exist: overall, plants that occur more frequently in a particular area seem to show a smaller degree of lexical diversity. However, the correlation is not strong enough for plant frequency to cause complete lexical uniformity within an ecological region and other factors play a role as well. Furthermore, a small-scale investigation of locally infrequent, but globally frequent plants revealed that the global frequency of a plant can cause a similar effect. However, more data is necessary to corroborate this finding.

Overall, we were able to show that the experiential environment of language users from different locations can contribute to explaining the amount of lexical variation a concept shows: varying environments and varying experiences do affect the language speakers use. Furthermore, we showed that using referential, rather than linguistic data to study the effect of the degree of experiential salience of a concept can provide further insight into lexical diversity.

Epilogue

7. Discussion

This dissertation examined the relationship between meaning and lexical diversity. The overarching aim was to contribute to lexical semantics from a Cognitive Sociolinguistics perspective by showing that lexical diversity is not only affected by socio-cultural properties of the dialects involved, but also by features related to the cognitive aspects of categorization, language processing and production. In practice, two immediate aims (1 and 2) and two broader research goals (3 and 4) were distinguished:

1. to establish that the results obtained in the pilot studies, viz. that lack of onomasiological salience, onomasiological vagueness and proneness to affect, influence the amount of lexical diversity a concept shows, are stable in other semantic fields and dialect areas
2. to elaborate on the results obtained in the pilot studies and distinguish additional concept features that can influence the amount of lexical diversity a concept shows
3. to show how, by combining different methods, a better picture of the structure of lexical diversity can be obtained
4. to elucidate that introducing a theoretical linguistic perspective to dialectological research on lexical variation can be beneficial for both disciplines

Section 7.1 provides a summary of the case studies. In 7.2, the four ways in which each case study contributes to the systematization of and elaboration on the pilot studies (aims 1 and 2) are discussed in some more detail. 7.3 provides an overview of how this dissertation meets the broader research goals (3 and 4) and it outlines some suggestions for future research. A brief conclusion is provided in 7.4.

7.1. SUMMARY OF THE CASE STUDIES

In this dissertation, four case studies were presented that examine the relationship between lexical diversity and meaning in different ways. In part 1, we showed that cognitive concept features, related to the maximalist view on meaning of the Cognitive Linguistics paradigm, affect the amount of variation a concept shows. The main aim of the first case study, outlined in **chapter 3**, was to confirm that concept features related to the prototype-theoretical organization of the lexicon influence the amount of variation a concept shows in other datasets than were used in the pilot studies. More specifically, we examined the influence of these features on six semantic fields (the human body, the house, celebration & entertainment, personality & feelings, family & sexuality and society, school & education) of the digitized databases of the Dictionaries of the Brabantic and Limburgish Dialects. Using linear regression analysis, the correlation between lack of onomasiological salience, vagueness and proneness to affect and an operationalization of lexical diversity that takes into account both the number of different types that are available and the degree to which these variants are heterogeneously scattered across each dialect area per concept, was confirmed. The analysis shows that the results of the pilot studies are the same in other dialectal data and in other semantic fields: less salient, more vague and affect-sensitive concepts are characterized by significantly more lexical diversity. Additionally, we elaborated on the results of the pilot studies by selecting semantic fields organized along two dimensions (viz. concreteness and universal versus society-related versus locally bound concepts). This indicates that these dimensions affect the relative impact of the concept features between semantic fields. More specifically, on average, locally-bound concepts show more lexical geographical heterogeneity, especially when they

are not salient. The effect of onomasiological vagueness is stronger for fields that, on average, have a higher amount of concrete concepts.

In the second case study, presented in **chapter 4**, we inquired into the effect of these concept features on the different aspects of lexical diversity in dialect data, viz. number of unique types, dispersion and lack of spread per concept. First, we constructed separate linear regression models for each aspect of lexical diversity and compared the relative importance of the concept features. Then, we examined the extent to which the concept features are significant solely due to the fact that the data are geographically stratified. More specifically, we controlled for this signal by using the residuals of a regression model that predicts the number of unique types on the basis of the geographical variation per concept. Subsequently, we constructed a second linear regression model that takes into account the influence of the concept features on the variance that remains. This second part of the analysis can be considered as an exploratory way of establishing whether cognitive concept features are solely relevant because the data are geographically stratified and, thus, whether it is possible that they also affect differently stratified lexical data. The results of the regression models indicate that the effect of the concept features is relevant for every aspect of lexical diversity. However, some predictors play a larger role for the number of unique types (viz. affect and, to a lesser extent, onomasiological vagueness), while onomasiological salience seems to influence the geographical spread of the lexical variants more. These results were also confirmed in the second part of the analysis. Consequently, we concluded that the influence of onomasiological salience is of a different type than the influence of onomasiological vagueness and affect. The latter variables cause highly geographically diversified areas where particular variants are used and a larger number of different types per concept and, thus, a larger degree of synonymy within smaller areas or even in a single location. Onomasiological salience, however, seems to predominantly affect the degree of homogeneity in the (geographically stratified) onomasiological profile of a concept. Although non-salient concepts also have highly diversified geographical variant areas, this does not mean that within smaller areas, a significantly larger amount of synonymous variants occurs as well. We explained these findings by arguing that vaguer and affect-sensitive concepts are more prone to innovation and demarcational differences, while for onomasiologically less salient concepts, these feature are less important. Instead, perhaps dialect users merely rely on other words that they do know, like hyperonymous, co-hyponymous, or perhaps hyponymous lexical items.

In part 2, we took into account the fact that concept-related features can differ between language users. More specifically, we examined the extent to which the experiential and usage-based nature of meaning is reflected in lexical diversity between dialect speakers from different locations. **Chapter 5** investigated how the interaction between semantic features and lectal differences is reflected in the types of lexical variants that are used in different locations. In practice, we inquired into the usage of non-native variants from three source languages (viz. French, Latin and German) in four semantic fields (society, school & education, personality & feelings, church & religion and clothing & personal hygiene) in the Brabantic and Limburgish dialect area. The results indicate that we find geographical and semantic structure in the lexical variants that are used: loanwords are not used at random. The data show clear differences between the source languages and between the semantic fields. On the one hand, border effects for French and German show up in every semantic field and are dependent on geographical closeness. On the other hand, within each source language, we also find geographical patterns that are determined by the semantic field to which the concept belongs. These patterns can only be explained by relying on differences in socio-cultural history. The use of a larger or smaller amount of loanwords from a particular source language, thus, depends on the interaction between the geographical location of a dialect speaker and meaning, as particular source languages mostly affect particular semantic fields. Although these results corroborate previous findings, this case study contributes to contact linguistic research in Cognitive Sociolinguistics as it is one of the first to simultaneously investigate differences between source languages and in different semantic fields on a large scale.

In **chapter 6**, finally, we examined to what extent an experience-based characteristic of a concept, viz. experiential salience, correlates with lexical diversity. We relied on extra-linguistic, objective frequency counts of plants that occur naturally in the northern part of Belgium. The analysis shows that experiential salience correlates with lexical diversity: the more frequently a plant occurs in the everyday environment of a language user, the smaller the amount of unique types that are available, given the amount of tokens. However, as indicated in the discussion, experiential salience alone does not cause complete uniformity or semasiological conventionalization in a particular (ecological) region. More specifically, although the number of unique types decreases for more experientially salient plants, this does not mean that every dialect speaker from a particular ecological region always selects the same name. Other features play a role as well, like dialect borders within an ecological region, or the folkloristic relevance of a particular plant. In sum,

experiential frequency alone cannot cause complete lexical homogeneity: dialect speakers not only need to encounter a plant frequently in their everyday environment, but they need to talk about it as well.

7.2 ATTAINING THE IMMEDIATE RESEARCH GOALS

The first and second aims of this dissertation are to systematize and elaborate on the results of the pilot studies. The four case studies contribute to these goals in different ways, which can be organized along three dimensions. Table 7.1 summarizes the way in which these dimensions take form in each case study.

First, in contrast with the pilot studies, we took into account **other datasets** than the WLD. In every case study, we also included data from the WBD and in chapter 6, data from the WVD was used as well. This allows us to show that the results obtained in the pilot studies are valid for other dialect areas than the Limburgish one.

Second, although the pilot studies only focused on the human body, a universal semantic field, in this dissertation a **variety of semantic fields** was examined. The first three case studies that were presented use data from locally-bound, society-related and universal semantic fields. Additionally, in the final case study, a natural semantic field, the field of plants, was included as well. By relying on such a varied set of semantic fields, we have obtained evidence for the fact that the effect of concept-related features is relevant for many different types of referents in the referential world, from artefacts, over social concepts, to universal concepts and concepts that occur naturally.

Finally, the **operationalization of lexical diversity** differs between the chapters. In chapter 3, lexical diversity is calculated by means of a composite measure of number of unique types and geographical fragmentation and in chapter 4, the effect of the cognitive concept features on these two aspects of lexical diversity in the dialect data is examined. In chapter 5, we investigate how lexical diversity is formally reflected in the types of names that are used. In chapter 6, finally, different operationalizations that occur throughout this dissertation or in other studies are compared.

7.3 CONTRIBUTIONS TO THE BROADER AIMS AND SUGGESTIONS FOR FUTURE RESEARCH

Two additional broader research goals were distinguished as well. First, we aimed to show how **combining different methods** results in a better picture of the lexicon. This goal was achieved in two ways. On the one hand, as explained above, **different operationalizations of the concept of lexical diversity** were used throughout this dissertation and three operationalizations were directly compared in chapter 6. Chapter 2 already indicated that using different measures delivers diverging results. More specifically, the influence of onomasiological salience seems to be of a different type than the effect of onomasiological vagueness and affect. Additionally, the direct comparison of different measures of lexical diversity in chapter 6 results in a conclusion that shows the importance of taking into account the social perspective. Although lexical diversity, in its simplest form, takes into account the number of unique types available per concept (possibly in relation to the number of tokens), additionally, another crucial aspect lies in examining the degree to which the unique types are conventionalized in

	chapter 3 (case study 1)	chapter 4 (case study 2)	chapter 5 (case study 3)	chapter 6 (case study 4)
dialect areas	WBD & WLD	WBD & WLD	WBD & WLD	WBD, WLD & WVD (northern part of Belgium)
semantic fields	6 semantic fields that are universal, society-related or locally bound	6 semantic fields that are universal, society-related or locally bound	4 semantic fields, 3 of which are society-related and 1 of which is universal	1 semantic field of natural referents
operationalization of lexical diversity	number of unique types, geographical fragmentation	number of unique types versus geographical fragmentation	formal aspects of lexical diversity	number of unique types, TTR, internal uniformity

TABLE 7.1
Contribution of each case study to the systematization of and elaboration on the pilot studies

a speech community. Chapter 6, more specifically, showed that a fewer number of unique types per concept does not necessarily entail that every dialect speaker uses the same name in the same situation (in this case: ecological region).

On the other hand, the methods used throughout this dissertation are also diverse as they take into account the **geographical stratification of the data** in different ways. In chapter 3, we include geographical fragmentation directly into the response variable. In chapter 4, we follow a similar methodology, but also inquire into the effect of the concept features on lexical diversity when the geographical signal is accounted for. In chapter 5, we use the geographical distribution of the lexical variants, as a predictor variable in Generalized Additive Mixed Models. The interaction between a dialect speaker's geographical position (longitude and latitude) and the semantic field to which a concept belongs, are shown to have significant effects. Finally, in chapter 6, geography was operationalized by dividing the dialect areas included in the analyses, into ecologically consistent ecological regions. The analyses were then conducted within each region. By comparing these different approaches to geographical variation, we corroborated that the spatial distribution of dialect data plays a large role. For instance, the use of a particular non-native variant is highly dependent on the geographical location of a speaker. However, at the same time, it showed that geography alone cannot explain lexical diversity: next to the fact that dialectal data is geographically stratified, there are other reasons why in different dialects, different names are used. For some concepts, predominantly the more onomasiologically salient ones, the chance that one lexical variant becomes conventionalized within a dialect area is significantly larger. For other concepts, which are prone to affect or onomasiological vagueness, hardly any lexical uniformity occurs. Additionally, experiential salience alone does not cause conventionalization of a single lexical variant within an ecological region.

The second broader aim, to show how **a dialectological case study can contribute to theoretical linguistics and vice versa** is exemplified in all four case studies. First, dialectological data offer an interesting starting point for a study into lexical diversity, because they are characterized by geographical stratification and by the availability of a large number of heteronyms. As a result, by using the databases of the WBD and WLD, we were able to take a big data approach to the study of lexical diversity. Furthermore, dialectological varieties also show other types of stratification. We predominantly took advantage of this latter point in part 2 (chapters 5 and 6), by including the fact that meaning-related features can differ between people living in the same dialect area. Finally, although providing a dialect geographical picture of the data that we use was not an direct goal of this dissertation, we did mention in several chapters that an interesting addition to the studies would be to examine the influence of the concept features in smaller dialect regions, because this would offer further insight into the way a certain variant becomes conventionalized. Additionally, in chapter 5, a description of the Limburgish and Brabantic dialects was included, at least concerning the extent to which non-native lexical variants differs in the locations throughout each dialect area.

However, some **open questions** follow from our analyses. First, we took different **approaches to measuring meaning**. In part 1, aspects of meaning were measured on the basis of the linguistic data in the database, while in part 2, external, experiential features were used. Consequently, we found that semantic differences that are stable across dialect speakers and semantic features that are culturally determined both affect lexical diversity. However, the relationship between these aspects remains underdetermined. While we have proposed in chapter 6 that experiential frequency alone does not suffice to cause standardization, we cannot identify the precise relationship between experiential and onomasiological salience on the basis of the analyses that were conducted. For instance, does a concept only become onomasiologically salient when it is also experientially salient (in the broadest sense)? Is it enough for people to just talk about a particular concept to make it onomasiologically salient, although it is not necessary relevant in their environment? Furthermore, the nature of the interplay between experiential salience and onomasiological salience has not yet been made explicit. Is the effect of onomasiological and experiential salience identical? It is probably the case that experiential salience influences the social relevance of a concept, but if people do not engage in conversations about the concept, there will be no standardization (see chapter 6). Onomasiological salience is a multi-faceted phenomenon, which is not only influenced by the perceptual experience of a dialect speaker. Instead, it is an organizational property that is influenced by the dialectal relationship between experiential and linguistic relevance. A similar view has been advocated in the framework of cultural relativity (e.g. Sinha & Jensen de López 2001): cultural practices determine language, and, following from this, linguistic frequency reinforces what is culturally relevant. Future research can be envisaged that explores this relationship further. For instance, different types of data (like psychometric, linguistic and extra-linguistic) can be combined to examine to which extent they correlate and how the interaction between the different sources of salience affects lexical diversity in an identical way.

A further elaboration on the results obtained in this dissertation is to examine the effect of concept features in **other varieties**. First, this will shed light on some of the problems that were encountered, predominantly in chapter 6, concerning the comparison between the dictionaries. In chapter 6, we directly linked concepts together, but it turned out that this was not unproblematic. More specifically, only a limited set of concepts are identical in all of the dictionaries and, if they can be directly linked, the processes by means of which lexical variants are distinguished are not always identical (recall for instance that in the WLD, the phonological variant bosbessen 'bilberry' is grouped under the lexeme bosbes, while in the WBD, related phonological variants like bosbeize, bosbeze and bosbieseme are grouped under seperate lexical variants). By investigating the effect of the concept features in differently stratified varieties, additional evidence for their relevance would be obtained.

Second, although in chapter 4, we conducted a highly exploratory analysis of the extent to which the results that are obtained depend on the fact that dialectological data is geographically stratified, an additional question is whether concept-related features also affect differently stratified varieties. This question is an important one, because describing the "variation of meaning" (Geeraerts, Kristiansen & Peirsman 2010: 6) lies at the core of the Cognitive Sociolinguistics paradigm. Variation of meaning concerns the fact that, as exemplified in this dissertation, meaning interacts with other types of variation (like formal or conceptual onomasiological variation): choices between lexical alternatives are not only governed by the fact that synonymous expressions exist, which are lectally stratified and depend on the speech situation. They can, for instance, also be influenced by conceptual choices. Recall, for instance, the BLUE JEANS example outlined in chapter 1: BLUE JEANS can be conceptualized as a pair of jeans or as a pair of trousers, which involves different conceptual choices. Consequently, the paradigm has to come to terms with the interplay between form, meaning and context (ibid.: 8). Given that, as was shown in this dissertation, aspects of meaning can themselves influence the proneness to variability a concept shows, concept features should be taken into account to arrive at a truly valid description of the structure of the lexicon. For instance, if in two varieties, say British and American English, for particular concepts, the same word is used, this may be related to the fact that the concept is highly salient and not vague, which increases the likelihood of the varieties relying on the same word. For other less salient and more vague concepts, more variation will probably show up. If the aim would then be to establish the linguistic distance between British and American English, more reliable results will be obtained if the model takes into account the fact that some concepts are in general more or less prone to variation (also see Speelman & Geeraerts 2008).

Following from this interaction between form, meaning and context, two other aspects should be taken into account in follow-up research. First, the data for the concepts used in this dissertation were elicited by means of questionnaires. As a result, the context for each concept is relatively stable and the **contextual variation** has not yet been examined. The influence of onomasiological vagueness, for instance, may be highly dependent on context, as semasiological vagueness often involves a form of construal or contextual specification (Tuggy 1993). Perhaps the high number of lexical variants that occur in the database for vaguer concepts is related to the fact that different dialect speakers focus on (or construe into focus) a different aspect of the vague concept to be named. Consequently, taking into account differences in **conceptual onomasiological variation** would help elucidate such processes further.

Additionally, although in chapter 5, we inquired into the formal reflection of lexical diversity by examining the distribution of non-native versus native variants, perhaps different types of conceptual onomasiological variation are lectally stratified as well. It may, for instance, be the case that in places where a particular concept (say, a plant like the ROSA RUBIGINOSA, a type of rose native to Europe) is more onomasiologically salient, this is reflected in the fact that the name that is most often used for the concept is the unique name for this category (viz. sweet briar), whereas in other places where the plant is less well-known, hyperonymous or co-hyponymous names (like rose or rose moss, a common name for a different species, respectively) may be used. Although in this dissertation, we only inquired into formal onomasiological diversity, the extent to which this type of variation interacts with conceptual variation should be elucidated.

A final avenue for future research is to examine to what extent the concept features investigated, correlate with the **speed of lexical change**, i.e. the speed with which a particular lexical item is replaced by a new variant. It is a plausible assumption that concepts that show less lexical diversity from a synchronic perspective also show more diachronic uniformity, but this diachronic perspective was not examined in this dissertation. On the one hand, research in the Cognitive Linguistics paradigm has established that semasiological vagueness can serve as a catalyst of language change (e.g. Geeraerts 1997: chapter 3, Sweetser 1990). Consequently, perhaps more onomasiologically vague concepts are more prone to lexical change as well, due to the fact that the lexical items used for such concepts are less entrenched and conventionalized. On the other hand, a

number of studies in evolutionary linguistics have argued that word frequency correlates with language change: more frequent words are replaced at a slower rate (e.g. Bochkarev, Solovyev & Wichmann 2014, Pagel, Atkinson & Meade 2007; cf. Zipf's (1949: 109-120) principle of economical specialization, which in part entails that the frequency of a word is inversely related to its age). If all the lexical variants for a particular concept are taken into account, word frequency can be considered as an operationalization of onomasiological salience. As a result, in the framework advocated in this dissertation, the interpretation of the results obtained in such studies takes a different form. More specifically, they can be interpreted as preliminary evidence that a higher degree of onomasiological salience leads to a higher degree of diachronic conventionalization of particular lexical variants and, therefore, a slower rate of lexical change. However, as the onomasiological perspective is missing from most of these studies and as variation within a speech community is not directly taken into account, further research that does include these perspectives, is necessary. One way to attain this goal would, for instance, be to complement the data used in this dissertation with data from a different time period.

7.3 TO CONCLUDE

The picture of the dialect lexicon that emerges in this dissertation is, of course, a geographically stratified one. However, we were able to show that this geographical stratification is influenced by concept-dependent features related to the prototype-theoretical organization of the lexicon (onomasiological vagueness and salience), to encyclopaedic features of meaning and to the semantic fields to which the concept belongs (proneness to affect, concreteness and locally boundedness/society-relatedness/universality) and to the everyday experience of the dialect user. By conducting our analyses on dialectological data, we hope to have contributed to the knowledge of the structure of the lexicon at large. To round off this dissertation, a quote from Weijnen seems appropriate:

> *"Nog verder reikt trouwens de betekenis van de dialectologie. Zij is wel bij uitstek geschikt gebleken om een aantal vragen van algemeen taalkundige aard, zoal niet tot een oplossing te brengen, dan toch van een geheel nieuwe zijde te belichten" (Weijnen 1975c [1958]: 2).*

> *"The importance of dialectology reaches even further. It has been shown to be pre-eminently suitable to shed new light on questions in general linguistics, or even provide them with a solution."*

References

Allan, K. & Burridge, K. (1988). Euphemism, dysphemism, and cross-varietal synonymy. *La Trobe Working Papers in Linguistics, 1*, 1-16.

Allan, K. & Burridge, K. (2006). *Forbidden Words: Taboo and the Censoring of Language*. Cambridge: Cambridge University Press.

Auer, P. (2005). Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. In Delbecque, N., Van der Auwera, J. & Geeraerts, D. (eds.). *Perspectives on Variation. Sociolinguistic, Historical, Comparative*. Berlin, Boston: De Gruyter Mouton. 7-42.

Backus, A. (2014). A usage-based approach to borrowability. In Zenner, E. & Kristiansen, G. (eds.). *New Perspectives on Lexical Borrowing: Onomasiological, Methodological and Phraseological Innovations* (Language Contact and Bilingualism 7). Boston: De Gruyter Mouton. 19-40.

Barrett, L. F. & Russell, J. A. (1999). Structure of current affect. *Current Directions in Psychological Science, 8*, 10-14.

Belemans, R & Goossens, J. (2000). Inleiding bij deel III van het WBD. *Woordenboek van de Brabantse Dialecten*. Assen: Van Gorcum. 11-68.

Berlin, B. (1972). Speculations on the growth of ethnobotanical nomenclature. *Language in Society, 1*(1), 51-86.

Berlin, B. (1978). Ethnobiological classification. In Rosch, E. & Lloyd, B. *Cognition and Categorization*. New York: Wiley. 9-26.

Berlin, B. & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley: University of California press.

Berlin, B., Breedlove, D. & Raven, P. (1973). General Principles of Classification and Nomenclature in Folk Biology. *American Anthropologist, 75*(1), 214-242.

Bivand, R., Pebesma, E.J. & Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*. New York: Springer.

Blancquaert, E. & Pée, W. (eds.). (1925-1982). *Reeks Nederlandse Dialectatlassen (RND)*. 16 volumes. Antwerpen/Malle.

Bloomfield, L. (1958). *Language* (6th edition). London: George Allen & Unwin.

Blumenthal-Dramé, A., Hanulíková, A. & Kortmann, B. (2017). *Perceptual Linguistic Salience: Modeling Causes and Consequences* (Research topic: Frontiers in Psychology). Lausanne: Frontiers media.

Boas, F. (1911). Introduction. In Boas, F. *Handbook of American Indian Languages. Part 1* (Smithsonian Institution. Bureau of American Ethnology. Bulletin 40) Washington: Government Printing Office. 1-83.

Bochkarev V., Solovyev V., Wichmann S. (2014). Universals versus historical contingencies in lexical evolution. *Journal of the Royal Society Interface, 11*, 20140841.

Bradley, M.M. & Lang, P.J. (1999). *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*. Technical Report C-1. Gainesville: The Center for Research in Psychophysiology, University of Florida.

Britain, D. (2002). Space and spatial diffusion. In Chambers, J.K., Trudgill, P. & Schilling-Estes, N. *The Handbook of Language Variation and Change*. Malden/Oxford: Blackwell. 603-637.

Britain, D. (2011). The heterogeneous homogenisation of dialects in England. *Taal & Tongval 63*(1), 43-60.

Brok, H. (1991). *Enkele Bloemnamen in de Nederlandse Dialekten: Etnobotanische Nomenclatuur in het Nederlandse Taalgebied* (Publikaties van het P. J. Meertens-instituut voor Dialectologie, Volkskunde en Naamkunde van de Koninklijke Nederlandse Akademie van Wetenschappen 18). Amsterdam: Meertens Instituut.

Brok, H. (2003). *Publicaties over Plantennamen in Nederland, Nederlandstalig België en Frans-Vlaanderen* (Werken van de Koninklijke Commissie voor Toponymie en Dialectologie. Vlaamse Afdeling 24). Tongeren: Michiels.

Brok, H. (2006). *Stinkend-juffertje en Duivelskruid: Volksnamen van Planten*. Amsterdam: Salome.

Bromhead, H. (2011). Ethnogeographical categories in English and Pitjantjatjara/Yankunytjatjara. *Language Sciences, 33*(1), 58-75.

Brown, C., Holman, E., Wichmann, S. & Velupillai, V. (2009). Automated classification of the world's languages: A description of the method and preliminary results. *STUF - Language Typology and Universals / Sprachtypologie und Universalienforschung, 61*(4), 285-308.

Brugman, C. & Lakoff, G. (1988). Cognitive topology and lexical networks. In Small, S.L., Cottrell, G.W. & Tanenhaus, M.K. (eds.). *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann. 477-508.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W. & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica, 150*, 80-84.

Bybee, J. & Hopper, P. (2001). *Frequency and the Emergence of Linguistic Structure* (Typological Studies in Language 45). Amsterdam: Benjamins.

Chambers, J. K. & Trudgill, P. (1980). *Dialectology* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University press.

Cornelissen, G. (2007). Uit de woordenschat van het Kerkraads: Duitse woorden. In Keulen, R., Van de Wijngaard, T., Crompvoets, H. & Walraven, F. *Riek van Klank: Inleiding in de Limburgse Dialecten* (Veldeke Taalstudies 2). Sittard: Veldeke Limburg. 60-65.

Cornips, L., Swanenberg, J., Heeringa, W. & De Vriend, F. (2016). The relationship between first language acquisition and dialect variation: Linking resources from distinct disciplines in a CLARIN-NL project. *Lingua, 178*, 32-45.

Crawley, M. (2007). *The R Book*. Chichester: Wiley.

Croft, W. (2009). Toward a social cognitive linguistics. In Evans, V. & Pourcel, S. (eds.). *New Directions in Cognitive Linguistics*. Amsterdam/Philadelphia: John Benjamins. 395-420.

Croft, W. & Cruse, A.D. (2004). *Cognitive Linguistics*. Cambridge University Press.

Cysouw, M. & Comrie, B. (2009). How varied typologically are the languages of Africa? In Botha, R. & Knight, C. (eds.). *The Cradle of Language. Volume 2: African Perspectives*. Oxford: Oxford University Press. 189-203.

Daan, J. (1969). Dialecten. In Daan, J. & Blok, D.P. *Van Randstad tot Landrand. Toelichting bij de Kaart: Dialecten en Naamkunde* (Bijdragen en Mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam 37). Amsterdam: Noord-Hollandsche uitgeversmaatschappij. 7-43.

Dabrowksa, E. (2015). Individual differences in grammatical knowledge. In Dabrowksa, E. & Divjak, D. *Handbook of Cognitive Linguistics* (Handbücher Zur Sprach- Und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 39). Berlin/ Boston: De Gruyter Mouton. 650-668.

Dabrowksa, E. & Divjak, D. (2015) *Handbook of Cognitive Linguistics* (Handbücher Zur Sprach- Und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 39). Berlin/ Boston: De Gruyter Mouton.

Daems, J., Heylen, K. & Geeraerts, D. (2015). Wat dragen we vandaag: Een hemd met blazer of een shirt met jasje? *Taal en Tongval, 67*(2), 307-342.

Daems J., Zenner, E. & Geeraerts, D. (2016). Lexicale homogeniteit en lexicale voorkeur in de Nederlandse woordenschat van emoties. *Tijdschrift voor Nederlandse Taal- en Letterkunde, 132* (4), 276-319.

Daller, H., Van Hout, R. & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals, *Applied Linguistics, 24*(2), 197-222.

Dancygier, B. (2017). *The Cambridge Handbook of Cognitive Linguistics*. Cambridge University Press.

De Pascale, S. (Forthcoming). Cultural models in contact: The case of regional varieties of Italian. In Zenner, E., Backus A. & Winter-Froemel E. (eds.). *Cognitive Contact Linguistics* (Cognitive Linguistics Research). Berlin: De Gruyter Mouton.

De Vriend, F & Swanenberg, J. (2006). D-kwadraat: Digitale databanken en digitaal gereedschap voor WBD en WLD. *Nederlandse Taalkunde, 11*(4), 366-372.

De Vriend, F., Swanenberg, J. & Van Hout, R. (2007). Dialectgebieden in Brabant. Geografische clustering op basis van de ruwe lexicale gegevens van het Woordenboek van de Brabantse Dialecten. *Taal en Tongval, Theme number 20*, 83-110.

De Vriend, F., Giesbers, C., Van Hout, R. & Ten Bosch, L. (2008). The Dutch-German border: Relating linguistic, geographic and social distances. *International Journal of Humanities and Arts Computing, 2*(1-2), 119-134.

De Vriend, F., Boves, L., Van den Heuvel, H., Van Hout, R., Kruijsen, J. & Swanenberg, J. (2006). A unified structure for Dutch dialect dictionary data. *Proceedings of The Fifth International Conference on Language Resources and Evaluation* (LREC 2006), Genoa, Italy.

De Vriendt, S. (2004). *Brussels* (Taal in stad en land 2). Tielt: Lannoo.

Dirven, R. & Verspoor, M. (2004). *Cognitive Exploration of Language and Linguistics* (Cognitive Linguistics in Practice 1). Amsterdam/Philadelphia: John Benjamins.

Divjak, D. & Caldwell-Harris, C.L. (2015). Frequency and entrenchment. In Dabrowksa, E. & Divjak, D. *Handbook of Cognitive Linguistics* (Handbücher Zur Sprach- Und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 39). Berlin/Boston: De Gruyter Mouton. 53-74.

Elliott, P. & Wartenberg, D. (2004). Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives, 112*(9), 998–1006.

Eyler, J. (2013). Commentary: Confronting unexpected results: Edmund Parkes reviews John Snow, *International Journal of Epidemiology, 42*(6), 1559–1562.

Evans, V. & Green, M. (2006). *Cognitive Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.

Field, A., Miles, J. & Field, Z. (2012). *Discovering Statistics Using R*. Los Angeles: Sage.

Fox, J. & Weisberg, S. (2011). Multivariate linear models in R. An appendix to *An R Companion to Applied Regression*. Thousand Oaks: Sage.

Geeraerts, D. (1986). Over woordverlies in lexicaal-semantische overgangsgebieden I. *Tijdschrift voor Nederlandse Taal- en Letterkunde, 102*(3), 161-186.

Geeraerts, D. (1987). Over woordverlies in lexicaal-semantische overgangsgebieden II. *Tijdschrift voor Nederlandse Taal- en Letterkunde, 103*(1), 37-51.

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics, 4*(3), 223-272.

Geeraerts, D. (1997). *Diachronic Prototype Semantics: A Contribution to Historical Lexicology* (Oxford Studies in Lexicography and Lexicology). Oxford: Clarendon.

Geeraerts, D. (2005). Lectal variation and empirical data in Cognitive Linguistics. In Ruiz de Mendoza Ibáñez, F. & Peña Cervel, S. (eds.). *Cognitive Linguistics. Internal Dynamics and Interdisciplinary Interactions*. Berlin/ New York: Mouton de Gruyter. 163-189.

Geeraerts, D. (2006). *Cognitive Linguistics: Basic Readings*. Berlin: De Gruyter Mouton.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Geeraerts, D. (2015). Sense individuation. In Riemer, N. (ed.), *The Routledge Handbook of Semantics*. London: Routledge. 233-247.

Geeraerts, D. (2016). Entrenchment as onomasiological salience. In Schmid, H. (ed.). *Entrenchment and the Psychology of Language Learning. How we Reorganize and Adapt Linguistic Knowledge*. Berlin, Boston: De Gruyter Mouton. 153-174.

Geeraerts, D. & Cuyckens, H. (2007). *The Oxford Handbook of Cognitive Lingustics*. Oxford: Oxford University Press.

Geeraerts, D. & Grondelaers, S. (1995). Looking back at anger. Cultural traditions and metaphorical patterns. In Taylor, J. & MacLaury, R.E. (eds.), *Language and the Cognitive Construal of the World* (Trends in Linguistics. Studies and Monographs 82). Berlin/New York: De Gruyter Mouton. 153-180.

Geeraerts, D. & Speelman, D. (2010). Heterodox concept features and onomasiological heterogeneity in dialects. In Geeraerts, D., Kristiansen, G. & Peirsman, Y. (eds.). *Advances in Cognitive Sociolinguistics* (Cognitive Linguistics Research 45). Berlin: De Gruyter Mouton. 23-39.

Geeraerts, D. & Van de Velde, H. (2013). Supra-regional characteristics of colloquial Dutch. In Hinskens, F. & Taeldeman, J. (eds.) *Language and Space. An International Handbook of Linguistic Variation. Volume 3: Dutch*. Berlin/Boston: De Gruyter Mouton. 532-555.

Geeraerts, D., Grondelaers, S. & Bakema, P. (1994). *The Structure of Lexical Variation: Meaning, Naming, and Context* (Cognitive Linguistics Research 5). Berlin: De Gruyter Mouton.

Geeraerts, D., Grondelaers, S. & Speelman, D. (1999). *Convergentie en Divergentie in de Nederlandse Woordenschat: Een Onderzoek naar Kleding- en Voetbaltermen*. Amsterdam: Meertens Instituut.

Geeraerts, D., Kristiansen, G. & Peirsman, Y. (2010). *Advances in Cognitive Sociolinguistics* (Cognitive Linguistics Research 45). Berlin: De Gruyter Mouton.

Geerts, G. & Heestermans, H. (1984). *Van Dale Groot Woordenboek der Nederlandse Taal* (11th edition). Utrecht: Van Dale Lexicografie.

Giesbers, C. (2008). *Dialecten op de Grens van Twee Talen: Een Dialectologisch en Sociolinguïstisch Onderzoek in het Kleverlands Dialectgebied*. PhD dissertation. Nijmegen: Radboud University.

Goebl, H. (1984). *Dialektometrische Studien. Anhand Italoromanischer, Rätoromanischer und Galloromanischer Sprachmaterialien aus AIS und ALF*. Tübingen: Niemeyer.

Goebl, H. (2006). Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing, 21*(4), 411-35.

Goebl, H. (2010). Dialectology and quantitative mapping. In Lameli, A., Kehrein, R. & Rabanus, S. (eds.) *Language and Space. An International Handbook of Linguistic Variation. Volume 2: Language Mapping*. Berlin: De Gruyter Mouton. 433-457, 2201-2212.

Goossens, J. (1964). Enkel- en veeltoepasselijkheid van betekenaars op de taalkaart. In *Taalgeografie en Semantiek. Lezingen Gehouden voor de Dialectencommissie der Koninklijke Nederlandse Academie van Wetenschappen op 27 December 1962 door dr. J. Goossens en dr. Jan van Bakel* (Bijdragen en Mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen ter Amsterdam XXVIII). Amsterdam: Noord-Hollandse uitgevers maatschappij. 3-27.

Goossens, J. (1972). *Inleiding tot de Nederlandse Dialectologie* (Handelingen van de Koninklijke Commissie voor Toponymie en Dialectologie 44). Tongeren: Michiels.

Gorman, A.M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology, 61*(1), 23-29.

Gries, S. (2015). Polysemy. In Dabrowksa, E. & Divjak, D. *Handbook of Cognitive Linguistics* (Handbücher Zur Sprach- Und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 39). Berlin/Boston: De Gruyter Mouton. 472-490.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing, 22*(3), 251-270.

Grieve, J. (2013). A statistical comparison of regional phonetic and lexical variation in American English. *Literary and Linguistic Computing, 28*(1), 82-107.

Grieve, J. (2017). Applied lectometry: Using multivariate spatial analysis to identify cultural regions. *ICLAVE 9, Extending the Scope of Lectometry*, Malaga, Spain, June 7, 2017.

Grieve, J., Speelman, D. & Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change, 23*(2), 193-221.

Grondelaers S. & Geeraerts, D. (1998) Vagueness as a euphemistic strategy. In Athanasiadou, A. & Tabakowska, E. (eds.). *Speaking of Emotions: Conceptualisation and Expression* (Cognitive Linguistics Research 10). Berlin: De Gruyter Mouton. 357-374.

Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Harder, P. (2003). The status of linguistic facts: Rethinking the relation between cognition, social institution and utterance from a functional point of view. *Mind & Language, 18*(1), 52-76.

Hargis, C. & Gickling, E. (1978). The function of imagery in word recognition development. *The Reading Teacher, 31*(8), 870-874.

Haspelmath, M.& Tadmor, U. (2009). *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wold.clld.org, Accessed on 19 July 2017).

Havermans, N. & Hooge, M. (2011). *Kerkpraktijk in België: Resultaten van de Zondagstelling in Oktober 2009. Rapport ten behoeve van de Belgische Bischoppenconferentie*. Leuven: KU Leuven, Centrum voor Politicologie.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences Using Levenshtein Distance*. PhD dissertation. Nijmegen: Radboud University.

Heeringa, W. & Nerbonne, J. (2006). De analyse van taalvariatie in het Nederlandse dialectgebied: methoden en resultaten op basis van lexicon en uitspraak. *Nederlandse Taalkunde, 11*(3), 218-257.

Held, L., Natário, I., Fento, S.E., Rue, H. & Becke, N. (2005). Towards joint disease mapping. *Statistical Methods in Medical Research, 14*(1), 61–82.

Heylen, K., Wielfaert, T., Speelman, D. & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua, 157*, 153-172.

Hock, H.H. & Joseph, B.D. (1996). *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics* (Trends in Linguistics. Studies and Monographs 93). Berlin: De Gruyter.

Hoppenbrouwers, C. & Hoppenbrouwers, G. (2001). *De Indeling van de Nederlandse Streektalen: Dialecten van 156 Steden en Dorpen Geklasseerd volgens de FFM*. Assen: Van Gorcum.

Jäger, G. (2013). Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change, 3*(2), 245-291.

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. *Language Learning, 63*, 87-106.

Jensen de López, K., Hayashi, M. & Sinha, C. (2005). Early shaping of spatial meaning in three languages and cultures: Linguistic or cultural relativity? *LACUS Forum, 31*, 379-88.

Johnson, M. (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination and Reason*. Chicago: University of Chicago Press.

Kemmer, S. & Barlow, M. (2000). Introduction: A usage-based conception of language. In Barlow, M. & Kemmer, S. (eds.). *Usage-based Models of Language*. Stanford: CSLI Publications. vii - xxviii.

Keuleers, E., Stevens, M., Mandera, P. & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology, 68*(8), 1-62.

Kövecses, Z. (1989). *Emotion Concepts*. New York: Springer.

Kövecses, Z. (2005). *Metaphor in Culture: Universality and Variation*. Cambridge: Cambridge University Press.

Kristiansen, G. & Dirven, R. (2008). *Cognitive Sociolinguistics* (Cognitive Linguistics Research 39). Berlin/New York: De Gruyter Mouton.

Kruijsen, J. (1990). Woordgeografie van ontleningen in een contactsituatie. *Taal en Tongval, 52*, 4-45.

Kruijsen, J. (1996). De Nijmeegse dialectlexicografische projecten. *Trefwoord, 11*, 93-107.

Kruijsen, J. (2001). Inleiding bij deel III van het WLD: De Algemene Woordenschat. *Woordenboek van de Limburgse Dialecten*. Assen: Van Gorcum. V-L.

Kruijsen, J. & Van Keymeulen, J. (1997). The Southern Dutch Dialect Dictionaries. *Lexikos, 7* (AFRILEX series 7), 207-228.

Labov, W. (1969). Contraction, deletion and inherent variability of the English copula. *Language, 45*(4), 715-762.

Labov, W. (2007). Transmission and diffusion. *Language, 83*(2), 344-387.

Labov, W. (2010). *Principles of Linguistic Change. Volume 3: Cognitive and Cultural Factors* (Language in Society 39). Chichester: Wiley-Blackwell.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago: University of Chicago Press.

Lakoff, G. (1990). The invariance hypothesis. *Cognitive Linguistics, 1*(1), 39-74.

Lakoff, G. & Johnson, M. (1980). *Metaphors we Live by.* Chicago: University of Chicago Press.

Langacker, R. (1987). *Foundations of Cognitive Grammar. Volume 1: Theoretical Prerequisites.* Stanford: Stanford University Press.

Langacker, R. (1988a). A usage-based model. In Rudzka-Ostyn, B. (ed.). *Topics in Cognitive Linguistics.* Amsterdam: John Benjamins. 127-161.

Langacker, R. (1988b). An overview of Cognitive Grammar. In Rudzka-Ostyn, B. (ed.). *Topics in Cognitive Linguistics.* Amsterdam: John Benjamins. 3-48.

Langacker, R. (2008). *Cognitive Grammar: A Basic Introduction.* Oxford: Oxford University Press.

Levshina, N. (2015). *How to do Linguistics with R: Data Exploration and Statistical Analysis.* Amsterdam: Benjamins.

Lewandowska-Tomaszczyk, B. (2007). Polysemy, prototypes and radial categories. In Geeraerts, D. & Cuyckens, H. (eds.). *The Oxford Handbook of Cognitive Linguistics.* Oxford: Oxford University Press. 139-169.

Lillo, A. (2009). Drunk: The definitive drinker's dictionary. *International Journal of Lexicography, 23*(2), 242-245.

Luyckx, K. & Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing, 26*(1), 35–55.

Lyons, J. (1977). *Semantics.* Cambridge: Cambridge University Press.

Majid, A. & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition, 130*(2), 266-270.

Margry, P.J. & Caspers, C. (2000). *Bedevaartplaatsen in Nederland: Limburg.* Amsterdam: Meertens Instituut.

McMahon, A. (1994). *Understanding Language Change.* Cambridge: Cambridge University press.

Mervis, C. & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*(1), 89-115.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A., De Schryver, M., De Winne, J. & Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4300 Dutch words. *Behavior Research Methods, 45*(1), 169-177.

Nederlandse Taalunie. (1990). Permanent overleg regionale woordenboeken. *Nederlandse Taalunie Publikatieblad, 3*(4).94-99.

Nerbonne, J & Kleiweg, P. (2003). Lexical distance in LAMSAS. *Computers and the Humanities, 37*(3), 339-57.

Nerbonne, J. & Kretzschmar, W. (2003). Introducing computational techniques in dialectometry. *Computers and the Humanities, 37*(3), 245-255.

Nichols, J. (2013). The vertical archipelago: Adding the third dimension to linguistic geography. In Auer, P., Hilpert, M., Stukenbrock, A. & Szmrecsanyi, B. (eds.). *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives* (Linguae & Litterae 24). Berlin: De Gruyter Mouton. 38-60.

Onysko, A. & Winter-Froemel, E. (2011). Necessary loans - luxury loans? Exploring the pragmatic dimension of borrowing. *Journal of Pragmatics, 43*(6), 1550-1567.

Osgood, C.E., Suci, G.J. & Tannenbaum, P.H. (1957). *The Measurement of Meaning.* Urbana: University of Illinois press.

Pagel, M., Atkinson, Q.D. & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature, 449*, 717-720

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach* (Oxford Psychology Series 9). New York: Oxford University Press.

Paul, H. (1891). *Principles of the History of Language. Translated from the Second Edition of the Original by H.A. Strong, M.A., L.L.D.* London: Longmans, Green and co.

Pauwels, J. (1933). *Enkele Bloemnamen in de Zuidnederlandsche Dialecten* (Noord- en Zuid-Nederlandsche Dialectbibliotheek 5). The Hague: Nijhoff.

Peersman, C., Rutten, G. & Vosters, R. (2015). *Past, Present and Future of a Language Border. Germanic-Romance Encounters in the Low Countries* (Language and Social Life 1). Berlin / New York: De Gruyter Mouton.

Philippa, M., Debrabandere, F., Quak, A., Schoonheim, T. & Van der Sijs, N. (2003-2009). *Etymologisch Woordenboek van het Nederlands.* Amsterdam: Amsterdam University Press

Pickl, S. (2013). Lexical meaning and spatial distribution. Evidence from geostatistical dialectometry. *Literary and Linguistic Computing, 28*(1), 63 -81.

Pizarro Pedraza, A. 2015. Who said "abortion"? Semantic variation and ideology in Spanish newspapers' online discussions. *Australian Journal of Linguistics, 35*(1), 53-75.

Preston, D.R. (1999). Introduction. In Preston, D.R. *Handbook of Perceptual Dialectology. Volume 1.* Amsterdam: Benjamins. xxiii-xl.

Pröll, S. (2013). Detecting structures in linguistic maps: Fuzzy clustering for pattern recognition in geostatistical dialectometry. *Literary and Linguistic Computing, 28*(1), 108-118.

Pütz, M., Robinson, J.A. & Reif, M. (2014). *Cognitive Sociolinguistics: Social and Cultural Variation in Cognition and Language Use* (Benjamins Current Topics 59). Amsterdam: John Benjamins Publishing Company.

Regier, T., Carstensen, A. & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PLoS ONE, 11*(4), e0151138.

Rohrer, T. (2007). Embodiment and experientialism. In Geeraerts, D. & Cuyckens, H. (eds.). *The Oxford Handbook of Cognitive Linguistics.* New York: Oxford University Press. 25-47.

Rosch, E. (1978). Principles of categorization. In Rosch, E. & Lloyd, B.B. (eds.), *Cognition and Categorization.* New York: Wiley. 27–48.

Rosch, E. (1987). Linguistic relativity. *ETC: A Review of General Semantics, 44*(3), 254-279.

Rosch, E. & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573-605

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Roukens, W. (1961). Uit de geschiedenis van het Kerkraads dialect. *Veldeke 37* (Kerkrade-nummer), 98-115.

Ruette, T. & Speelman, D. (2013). Transparent aggregation of variables with Individual Differences Scaling. *Literary and Linguistic Computing, 29*(1), 89-106.

Ruette, T. Ehret, K. & Szmrecsanyi, B. (2016). A lectometric analysis of aggregated lexical variation in written Standard English with Semantic Vector Space models. *International Journal of Corpus Linguistics, 21*(1), 48-79.

Ruette, T., Speelman, D. & Geeraerts, D. (2014). Lexical variation in aggregate perspective. In Soares da Silva A. (ed.), *Pluricentricity: Language Variation and Sociocognitive Dimensions* (Applications of Cognitive Linguistics 24). Berlin: Mouton de Gruyter. 103-126.

Rumpf, J., Pickl, S., Elspass, S., König, W. & Scmidt, V. (2010). Quantification and statistical analysis of structural similarities in dialectological area-class maps. *Dialectologia et Geolinguistica, 18*, 73-100.

Salverda de Grave, J.J. (1920). *De Franse Woorden in het Nederlands* (Verhandelingen der Koninklijke Akademie van Wetenschappen te Amsterdam. Afdeeling Letterkunde. Nieuwe Reeks VII). Amsterdam: Johannes Müller.

Schmeets, H. (2014). De *Religieuze Kaart van Nederland, 2010-2013.* s.l.: Centraal Bureau voor de Statistiek.

Schmid, H. (2015). A blueprint of the Entrenchment-and-Conventionalization Model. *Yearbook of the German Cognitive Linguistics Association, 3*(1), 3-26.

Schmid, H. (2016a). Linguistic entrenchment and its psychological foundations. In Schmid, H. (ed.). *Entrenchment and the Psychology of Language Learning. How we Reorganize and Adapt Linguistic Knowledge* (Language and the Human Lifespan 4). Berlin, Boston: De Gruyter Mouton. 435-452.

Schmid, H. (2016b). Why Cognitive Linguistics must embrace the social and pragmatic dimensions of language and how it could do so more seriously. *Cognitive Linguistics, 27*(4), 543-557.

Schuchardt, H. (1912). Sachen und Wörter. *Anthropos, 7*(4), 827-839.

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane, 35*, 335-357.

Sevenant, M., Menschaert, J., Couvreur, M., Ronse, A., Heyn, M., Janssen, J., Antrop, M., Geypens, M., Hermy, M. & De Blust, G. (2002). *Ecodistricten: Ruimtelijke Eenheden voor Gebiedsgericht Milieubeleid in Vlaanderen. Studieopdracht in het kader van Actie 134 van het Vlaams Milieubeleidsplan 1997-2001.* Commissioned by the Ministry of the Flemish Community. Administration of the Environment, Nature and Land Use.

Sinha, C. (1999). Grounding, mapping and acts of meaning. In Janssen, J. & Redeker, G. (eds.). *Cognitive Linguistics: Foundations, Scope and Methodology*. Berlin / New York: De Gruyter Mouton. 223-255.

Sinha, C. & Jensen de López, K. (2001). Language, culture and the embodiment of spatial cognition. *Cognitive Linguistics, 11*(1-2), 17-41.

Speelman, D. & Geeraerts, D. (2007). De structuur van lexicale onzekerheid. *Taal en Tongval, Theme number 20*, 47-61.

Speelman, D. & Geeraerts, D. (2008). The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing, 2*(1-2), 221-242.

Speelman, D. & Heylen, K. (2017). From dialectometry to semantics. In Wieling, M., Kroon, M., Van Noord, G. & Bouma, G. (eds.) *From Semantics to Dialectometry: Festschrift in honor of John Nerbonne* (Tributes 32). 325-334.

Speelman, D., Grondelaers, S. & Geeraerts, D. (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities, 37*(3), 317-37.

Swanenberg, J. (2000). *Lexicale Variatie Cognitief-semantisch Benaderd: Over het Benoemen van Vogels in Zuid-Nederlandse dialecten*. PhD dissertation. Nijmegen: Radboud University.

Swanenberg, J. (2004). Kolbloemen, bloedpaters en manebladen versus zoetelief en luitentuit. Bronnen van lexicale variatie in de Brabantse flora en fauna. *Taal & Tongval, 56*, 19-47.

Swanenberg, J. (2010). Als het beestje maar een naam heeft. In Rooryck, J., Van Craenenbroeck, J., Vanden Wyngaerd, G., De Caluwe, J. & Van Keymeulen, J. *Voor Magda: Artikelen voor Magda Devos bij haar Afscheid van de Universiteit Gent*. Ghent: Academia press. 561-568.

Sweetser, E. (1990). *From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure*. (Cambridge Studies in Linguistics 54). Cambridge: Cambridge University Press.

Szelid, V. & Geeraerts, D. (2008). Usage-based dialectology. Emotion concepts in the Southern Csango dialect. *Annual Review of Cognitive Linguistics*, 6, 23-49.

Szmrecsanyi, B. (2008). Corpus-based dialectometry: Aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing, 2*(1-2), 279-296.

Szmrecsanyi, B (2013). *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry* (Studies in English Language). Cambridge: Cambridge University Press.

Szmrecsanyi, B. (2016). About text frequencies in historical linguistics: Disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory, 12*(1), 153-171.

Tadmor, U. (2009). Loanwords in the world's languages: findings and results. In Haspelmath, M. & Tadmor, U. *Loanwords in the World's Languages*. Berlin / Boston: De Gruyter Mouton. 55-75.

Talmy, L. (1978). The relation of grammar to cognition – a synopsis. In Waltz, D. *Proceedings of TINLAP 2 (Theoretical Issues in Natural Language Processing)*. New York: Association for Computing Machinery. 14-24.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In Shopen, T. (ed.), *Language Typology and Syntactic Description. Volume 3: Grammatical Categories and the Lexicon*. Cambridge: Cambridge University Press. 57-149.

Talmy, L. (2006). Grammatical construal. The relation of grammar to cognition. In Geeraerts, D. (ed.) *Cognitive Linguistics: Basic Readings*. Berlin: De Gruyter Mouton. 69-108.

Theissen, S. (1975). *De Germanismen in de Moderne Nederlandse Woordenschat* (Bouwstoffen en Studiën voor de Geschiedenis en de Lexicografie van het Nederlands XIII). s.l.: Belgisch Interuniversitair Centrum voor Neerlandistiek.

Thomason, S. (2001). *Language Contact*. Edinburgh: Edinburgh university press.

Tomasello, M. (2001). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics, 11*, 61–82.

Tuggy, David. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics, 4*, 273–90.

Tweedie, F.J. & Baayen, H.R. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*, 323–352.

Van de Wijngaard, T. (2007). De Ripuarische dialecten. In Keulen, R., Van de Wijngaard, T., Crompvoets, H. & Walraven, F, *Riek van Klank: Inleiding in de Limburgse Dialecten* (Veldeke Taalstudies 2). Sittard: Veldeke Limburg. 45-59.

Van de Wijngaard, T. & Keulen, R. (2007). De indeling van de Limburgse dialecten. In Keulen, R., Van de Wijngaard, T., Crompvoets, H. & Walraven, F, *Riek van Klank: Inleiding in de Limburgse Dialecten* (Veldeke Taalstudies 2). Sittard: Veldeke Limburg. 15-23.

Van der Sijs, N. (2005). *Groot Leenwoordenboek*. Utrecht: Van Dale Lexicografie.

Van der Sijs, N. & Engelsman, J. (2000). *Nota Bene: De Invloed van het Latijn en Grieks op het Nederlands*. 's-Gravenhage: SDU.

Van Hout, R., Kruijsen, J. & Gerritsen, M. (2014). Exosmosis along the Romance-Germanic language border in Belgium: The diffusion of French borrowings in the Dutch dialects of Haspengouw. In Casesnoves-Ferrer, R., Forcadell Guinjoan, M., Gavaldà Ferré, N. (eds.). *Ens Queda la Paraula: Estudis de Lingüística Aplicada en Honor a M. Teresa Turell*. Barcelona: Institut Universitari de Lingüística Aplicada. 197-220.

Van Keymeulen, J. (1992). *De Algemene Woordenschat in de Grote Dialectwoordenboeken (WBD, WLD, WVD)*. PhD dissertation. Ghent: University of Ghent.

Van Landuyt, W., Hoste, I., Vanhecke, L., Van den Bremt, P., Vercruysse, W. & De Beer, D. (2006). *Atlas van de Flora van Vlaanderen en het Brussels Gewest*. Brussels: Research Institute for Nature and Forest, National Botanic Garden of Belgium & Flo.Wer.

Van Rij, J. (2015). *Overview GAMM Analysis of Time-series Data*. (Available online at http://www.sfs.uni-tuebingen.de/~jvanrij/Tutorial/GAMM.html, Accessed on 4 July 2017.)

WBD = *Woordenboek van de Brabantse Dialecten*. (1967-2005). Assen: Van Gorcum / Amsterdam: Gopher.

Weijnen, A. (1946). De grenzen tussen de Oost-Noord-Brabantse dialecten onderling. In *Oost-Noord-Brabantse Dialectproblemen. Lezingen Gehouden voor de Dialectencommissie der Koninklijke Nederlandse Akademie van Wetenschappen op 12 april 1944*. (Bijdragen en Mededelingen van de Dialectencommissie 8). Amsterdam: Noord-Hollandsche Uitgeverij Maatschappij. 1-17.

Weijnen, A. (1966). *Nederlandse Dialectkunde* (Studia Theodisca 10). Assen: Van Gorcum.

Weijnen, A. (1967). Leenwoorden uit de Latinitas stratigrafisch beschouwd. *Verslagen en Mededelingen van de Koninklijke Academie voor Nederlandse Taal- en Letterkunde, 5*, 365-480.

Weijnen, A. (1975a). De semantische en syntactische problematiek van het dialectwoordenboek. In Weijnen, A. *Algemene en Vergelijkende Dialectologie / General and Comparative Dialectology*. Amsterdam: Holland University Press. 23-33.

Weijnen, A. (1975b). De waarde van een dialectwoordenboek. In Weijnen, A. *Algemene en Vergelijkende Dialectologie / General and Comparative Dialectology*. Amsterdam: Holland University Press. 83-90.

Weijnen, A. (1975c). De oriëntatie van de dialectstudie. In Weijnen, A. *Algemene en Vergelijkende Dialectologie / General and Comparative Dialectology*. Amsterdam: Holland University Press. 1-15.

Weijnen, A. & Van Bakel, J. (1967). *Voorlopige Inleiding op het Woordenboek van de Brabantse Dialecten*. Assen: Van Gorcum.

Weijnen, A., Goossens, J. & Goossens, P. (1983). Inleiding. In Weijnen, A., Goossens, J. & Goossens P. (eds.) *Woordenboek van de Limburgse Dialecten: Inleiding & I. Agrarische Terminologie, Aflevering 1: Akker- en Weidegrond*. Assen: Van Gorcum. 1-77.

Weinreich, U. (1968). *Languages in Contact: Findings and Problems*. The Hague: De Gruyter Mouton.

Weinreich, U., Labov, W. & Herzog, M. (1968). Empirical foundations for a theory of language change. In Lehmann, W., Malkiel, Y. (eds.). *Directions for Historical Linguistics*. Austin: University of Texas Press. 95-195.

Whorf, B. (1964). Science and linguistics. In Whorf, B. & Carroll, J.B. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (MIT Press Paperback Series 5). Cambridge: MIT Press. 207-219.

Wieling, M. (2012). *A Quantitative Approach to Social and Geographical Dialect Variation*. PhD dissertation. Groningen: University of Groningen.

Wieling, M. (2017). *Statistics Course on Generalized Additive Modeling* (Available online at http://www.let.rug.nl/~wieling/statscourse/, Accessed on 4 July 2017).

Wieling, M., Nerbonne, J. & Baayen, R. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE, 6*(9), E23613.

Wieling, M., Upton, C. & Thompson, A. (2014). Analyzing the BBC Voices data: Contemporary English dialect areas and their characteristic lexical variants. *Literary and Linguistic Computing, 29*(1), 107-117.

WLD = *Woordenboek van de Limburgse Dialecten*. (1983-2008). Assen: Van Gorcum / Amsterdam: Gopher.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R* (Texts in Statistical Science). Boca Raton: Chapman and Hall/CRC.

Wood, S. (2017). *Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. (Available online at https://cran.r-project.org/web/packages/mgcv/mgcv.pdf, Accessed on 15 March 2017).

WVD = *Woordenboek van de Vlaamse Dialecten*. (1972-). Ghent: Academia Press.

Zenner, E. (2013). *Cognitive Contact Linguistics: The Macro, Meso and Micro Influence of English on Dutch*. PhD dissertation. Leuven: KU Leuven.

Zenner, E. & Kristiansen, G. (2014). Introduction: Onomasiological, methodological and phraseological perspectives on lexical borrowing. In Zenner, E. & Kristiansen, G. (eds.). *New Perspectives on Lexical Borrowing: Onomasiological, Methodological and Phraseological Innovations* (Language Contact and Bilingualism 7). Boston: De Gruyter Mouton. 1-17.

Zenner, E., Speelman, D. & Geeraerts, D. (2012). Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics, 23*(4), 749-792.

Zenner, E., Speelman, D. & Geeraerts, D. (2014). Core vocabulary, borrowability, and entrenchment: A usage-based onomasiological approach. *Diachronica, 31*(1), 74-105.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Cambridge: Addison-Wesley Press.

Zlatev, J. (1997). *Situated Embodiment: Studies in the Emergence of Spatial Meaning*. Stockholm: Gotab.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A. & Smith, G.M. (2009). *Mixed Effects Models and Extensions in Ecology with R* (Statistics for Biology and Health). New York: Springer.

List of Figures

List of Tables

Appendices

CHAPTER 3

3.1

Overview of semantic subdomains and example concepts per semantic field in the WLD

the human body	
subsection	**examples**
the body & body parts	bones, upperpart of the back, bristly hair, index finger
the senses	to spy on someone, to hear, to smell, flavour, (to be) numb
the organs	to breathe, blood, intestines, genitalia, nerve
the house	
subsection	**examples**
in general	country house, to move
parts of the building	lightning rod, tile, landing, door handle, key
rooms inside the house	corridor, restroom, attic room
the kitchen	furnace, pantry, knife to cut bread with, skimmer
the bathroom	plug (of a bathtub)
furniture	blanket, mattress, sideboard, baby chair
upholstery	net curtain, rug, stair rod
kitchenware	preserving jar, Cologne pot, jug
service, cutlery, glassware	milk jug, to set the table, glass bell, basket
upkeep, dishes and laundry	to tidy up, to do the dishes, mangle, shoe brush, to sweep
other activities inside the house	to embroider, to take a nap, to stew
electricity	fuse
lighting	sconce, chandelier, hanging lamp, to put the light on
heating	wood chip, soot, to fume, chimney
to make fire	tinderbox, firestone, sulphur
plants in house and garden	chrysanthemum, fuchsia, petunia, pear tree
garden	lawn, fencing, terrace
to work in the garden	rake, to mow the lawn, to do the weeding

the house	
pets	to bark, male dog, young kitten, nickname for a cat
celebration & entertainment	
subsection	**examples**
festivities and customs	name day, Shrove Tuesday, to visit a new mother and her baby, New Year, money to go to the fair, to look for Easter eggs, tree used for topping-out ceremony
sports and games	cheater, to play shuffleboard, to play marbles, cycle race
the arts	marble statue, trumpet, theatre
personality & feelings	
subsection	**examples**
intellectual capacity and memory	thinking, knowing, smart, dumb, to judge/to consider
personality	(un)reliable, (in)sincere, diligent-lazy, brave-frightened
feelings	fun, laughter, anger, sadness, disappointment
behaviour	to behave, dominance, to (dis)obey, success-failure, (in)decency
family & sexuality	
subsection	**examples**
human life in general	young woman, to give birth, death
descent, family and kinship	descent, family, kinship
birth and baptism	child to be baptized, godfather, godchild
dating, engagement, marriage	girlfriend, to match, bride, to live together without being married
sexual life	bastard child, mistress, prostitute, sexual intercourse
death and burial	coffin, funeral, widow's veil, guardian
society, school & education	
subsection	**examples**
man and society	labour, money, gift, to love, to like someone, cordial, to complain, to lie, news, postman
societal organization	marketplace, governor, police, perjury, war, border
school and education	boarding school, teacher, ruler, principle
transportation	car, rails, canal, hot air balloon, passport, to travel

CHAPTER 5

5.1 Results of the GA(M)Ms: maps with differing limits for the colour schemes per semantic field

5.1.1 French tokens

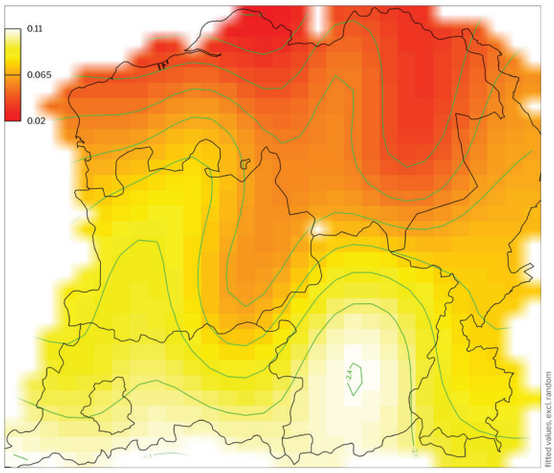


FIGURE A5.1.1A
French tokens per location
(society, school & education)

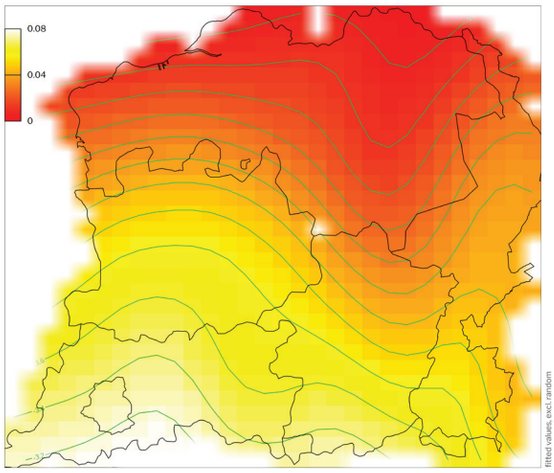


FIGURE A5.1.1B
French tokens per location
(personality & feelings)

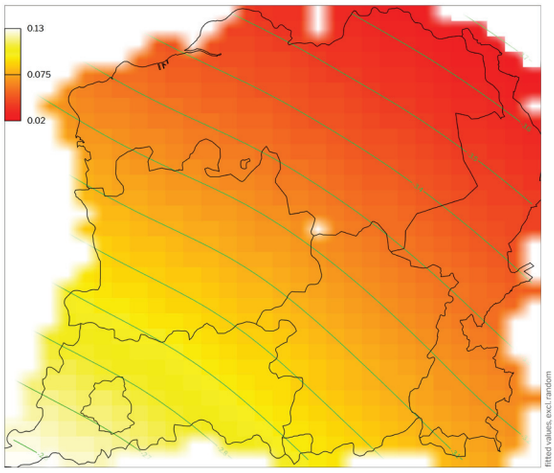


FIGURE A5.1.1C
French tokens per location
(church & religion)

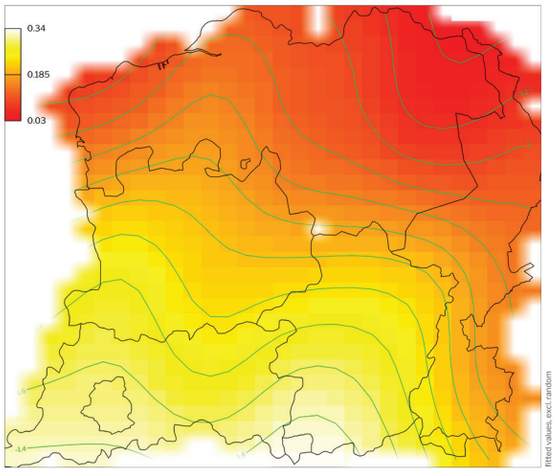


FIGURE A5.1.1D
French tokens per location
(clothing & personal hygiene)

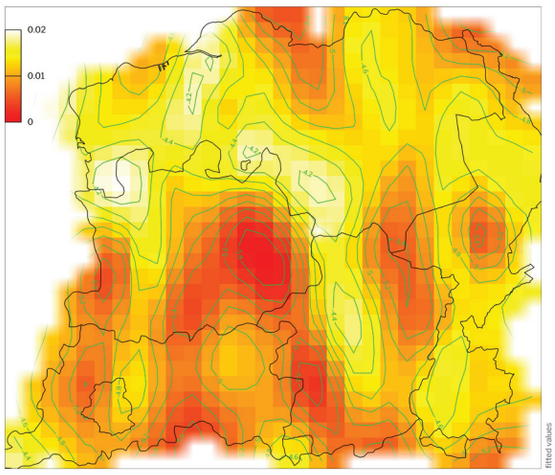


FIGURE A5.1.2A
Latin tokens per location
(society, school & education)

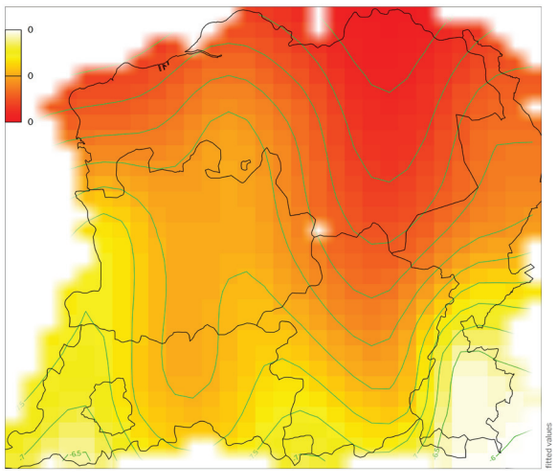


FIGURE A5.1.2B
Latin tokens per location
(personality & feelings)

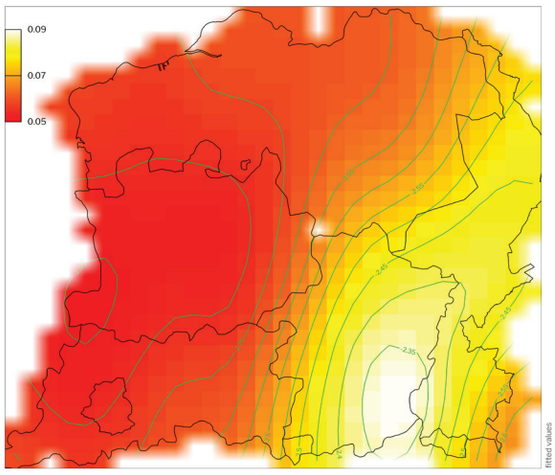


FIGURE A5.1.2C
Latin tokens per location
(church & religion)

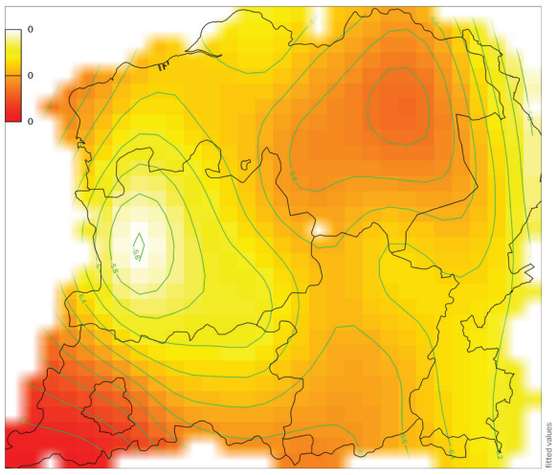


FIGURE A5.1.2D
Latin tokens per location
(clothing & personal hygiene)

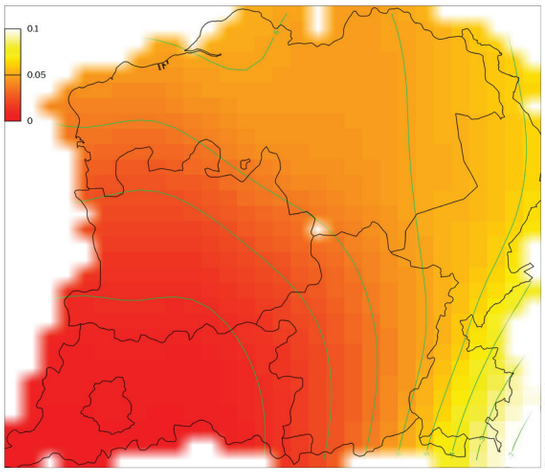


FIGURE A5.1.3A
German tokens per location
(society, school & education)

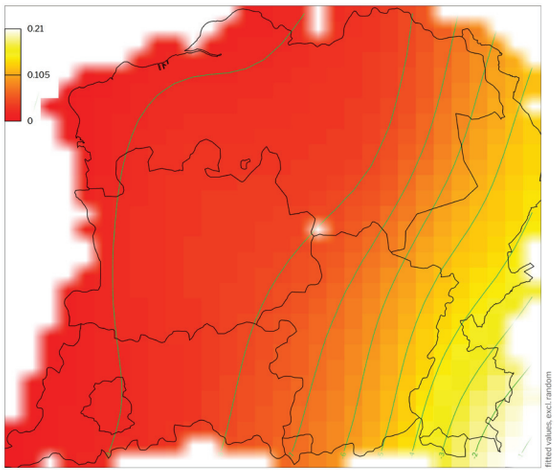


FIGURE A5.1.3B
German tokens per location
(personality & feelings)

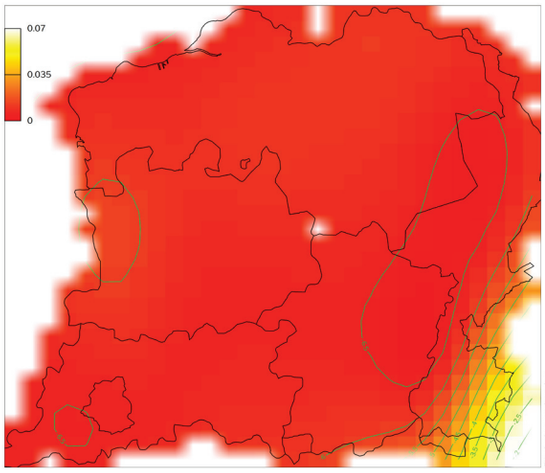


FIGURE A5.1.3C
German tokens per location
(church & religion)

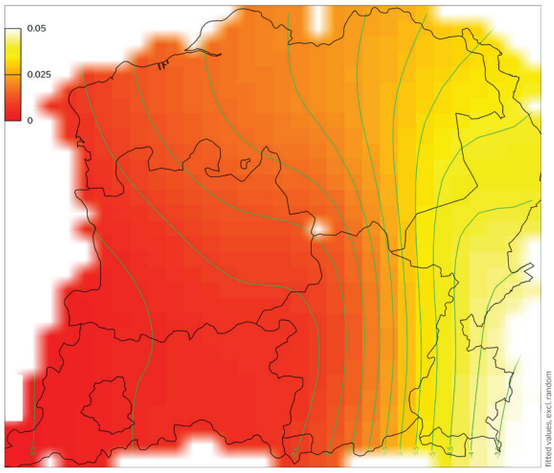


FIGURE A5.1.3D
German tokens per location
(clothing & personal hygiene)

5.2
Distribution of "German" concepts that occur in at least two locations in the Ripuarian dialect area in the semantic field of personality and feelings

concept	translation	nr. of types	nr. of observations	nr. of German types	nr. of German obs.	prop. of German types	prop. of German tokens
onnozel persoon	nitwit	9	10	1	1	0.111	0.100
snotneus	brat	5	10	1	1	0.200	0.100
prutsen	to mess about	6	6	1	1	0.167	0.167
begrip, besef	understanding	3	5	1	1	0.333	0.200
mokken	to sulk	5	5	1	1	0.200	0.200
bezorgd	concerned	3	4	1	1	0.333	0.250
kouwe drukte	much ado about nothing	4	4	1	1	0.250	0.250
mopperen	to grumble	4	4	1	1	0.250	0.250
potsachtig	comical	4	4	1	1	0.250	0.250
stiekem	sneaky	4	8	2	2	0.500	0.250
teleurgesteld (worden)	(to be) disappointed	3	4	1	1	0.333	0.250
tevreden, tevredenheid	satisfied, satisfaction	3	4	1	1	0.333	0.250
beestachtig persoon, beestachtig	savage (person)	3	3	1	1	0.333	0.333
beteuterd	dismayed	3	3	1	1	0.333	0.333
de baas spelen	to boss some-one around	3	3	1	1	0.333	0.333
genoegen (doen)	(to be) satisfactory	2	3	1	1	0.500	0.333
gluiperd	shifty character	3	3	1	1	0.333	0.333
lasteren	to insult	3	3	1	1	0.333	0.333
onwennig (voelen)	(to feel) ill at ease	3	3	1	1	0.333	0.333
treuren	to be sorrowful	3	3	1	1	0.333	0.333
van katoen geven	to give it all one has got	3	3	1	1	0.333	0.333
zich vergissen	to be mistaken	3	3	1	1	0.333	0.333
zich zeer slecht gedragen	to behave very badly	2	3	1	1	0.500	0.333
bangerik	coward	5	9	1	4	0.200	0.444

concept	translation	nr. of types	nr. of observations	nr. of German types	nr. of German obs.	prop. of German types	prop. of German tokens
aandringen	to insist	2	2	1	1	0.500	0.500
aanstoot geven	to give offence	2	2	1	1	0.500	0.500
aarzelen	to hesitate	2	2	1	1	0.500	0.500
angst	fear	2	2	1	1	0.500	0.500
baldadig (persoon)	rowdy (person)	2	2	1	1	0.500	0.500
bedrieger	fraud	2	2	1	1	0.500	0.500
bestemmen	to reserve	2	2	1	1	0.500	0.500
bezadigd	steady	2	2	1	1	0.500	0.500
boertig	coarse	2	2	1	1	0.500	0.500
geheimzinnig	mysterious	2	2	1	1	0.500	0.500
gemakkelijk	easy	2	2	1	1	0.500	0.500
grapjas	joker	4	4	2	2	0.500	0.500
gril	whim	2	2	1	1	0.500	0.500
hansworst	buffoon	2	2	1	1	0.500	0.500
hopen	to hope	2	2	1	1	0.500	0.500
hulp, bijstand	help	2	2	1	1	0.500	0.500
iemand kwaad maken	to anger someone	2	2	1	1	0.500	0.500
ingetogen	modest	2	2	1	1	0.500	0.500
jaloers	jealous	2	2	1	1	0.500	0.500
kalm, bedaard	calm	2	2	1	1	0.500	0.500
keus	choice	2	2	1	1	0.500	0.500
kiezen	to choose	2	2	1	1	0.500	0.500
kniezen	to mope	2	2	1	1	0.500	0.500
knoeier	sloppy person	2	2	1	1	0.500	0.500
konkelen	to scheme	2	2	1	1	0.500	0.500
lichtgeraakt, kregel	touchy	2	2	1	1	0.500	0.500
lui	lazy	2	2	1	1	0.500	0.500
nauwgezet, nauwgezet persoon	conscientious (person)	2	2	1	1	0.500	0.500

concept	translation	nr. of types	nr. of observations	nr. of German types	nr. of German obs.	prop. of German types	prop. of German tokens
niet helder van geest	not lucid	2	2	1	1	0.500	0.500
schuchter	shy	2	2	1	1	0.500	0.500
slecht mens, slechte kerel	bad person, bad fellow	2	2	1	1	0.500	0.500
slordig	sloppy	2	2	1	1	0.500	0.500
troosten, troost	(to) comfort	2	2	1	1	0.500	0.500
verstandig	sensible	2	2	1	1	0.500	0.500
verzuimen	to neglect	2	2	1	1	0.500	0.500
zich gedragen	to behave	2	2	1	1	0.500	0.500
zich kwaad maken	to get angry	2	2	1	1	0.500	0.500
zonder opzet	unintentionally	2	2	1	1	0.500	0.500
zwoegen	to labour	4	4	2	2	0.500	0.500
slim	smart	4	7	1	4	0.250	0.571
treiteren	to torment	3	5	1	3	0.333	0.600
geestig	witty	3	3	2	2	0.667	0.667
huichelaar	hypocrite	3	3	2	2	0.667	0.667
iemand prijzen	to commend someone	3	3	2	2	0.667	0.667
informeren (onoverg.)	to inform	2	3	1	2	0.500	0.667
leep, doortrapt	cunning	3	3	2	2	0.667	0.667
prettig	pleasant	3	3	2	2	0.667	0.667
schelm	crook	3	3	2	2	0.667	0.667
uitbrander	dressing-down	2	3	1	2	0.500	0.667
vanzelf-sprekend	obvious	2	3	1	2	0.500	0.667
vermaak	entertainment	3	3	2	2	0.667	0.667
wijs	wise	2	3	1	2	0.500	0.667
zich bedenken	to change one's mind	3	3	2	2	0.667	0.667
begrijpen	to understand	2	4	1	3	0.500	0.750
ellende (lijden)	misery	3	4	2	3	0.667	0.750

concept	translation	nr. of types	nr. of observations	nr. of German types	nr. of German obs.	prop. of German types	prop. of German tokens
plezier maken	to have fun	2	4	1	3	0.500	0.750
pret, schik	fun	3	4	2	3	0.667	0.750
pretmaker	merrymaker	3	4	2	3	0.667	0.750
vrolijk	cheerful	4	6	3	5	0.750	0.833
degelijk	decent	1	2	1	2	1.000	1.000
dwingen	to force	1	2	1	2	1.000	1.000
eenvoudig	simple	1	2	1	2	1.000	1.000
gunst	favor	2	2	2	2	1.000	1.000
ophouden met het werk	to end the working day	2	3	2	3	1.000	1.000
schipperen	to compromise	2	2	2	2	1.000	1.000
slimmerik	smart number	3	4	3	4	1.000	1.000
sober	sober	1	2	1	2	1.000	1.000
vreugde	joy	2	2	2	2	1.000	1.000
zedig	chaste	1	2	1	2	1.000	1.000
zijn tevreden-heid betuigen	to express one's satisfaction	2	3	2	3	1.000	1.000

6.1 Lexical items for plants in Table 6.6

lexical item	N	lexical item	N	lexical item	N
gele kaars	1	toorts	1	zoklappen	1
gele thee	1	toppen	1	kalverwortel	1
kattenkop	1	wilde zokken	1	kaars	2
koningskaars	1	wolplant	1	paaskaars	2
lammetjesblaren	1	wolvenstaart	1	wilde tabak	2
lammetjesoren	1	zokjes	1	wolharen	2
maagdenkaars	1	zokken	1		
stalkaars	1	zokkenblaren	1		

6.1.1
Distribution of lexemes for the great mullein (Verbascum Thapsus) in the Loamy region

lexical item	N	lexical item	N
wilde zuring	1	schape-, schaap(s)zurkel	4
dokke	1	wilde zurkel	10
paardezurkel	3	zurkel	19

6.1.2
Distribution of lexemes for the bitter dock (Rumex obtusifolius) in the Polder region

lexical item	N
acajou	1
robinia	1
valse acacia	1
acacia	23

6.1.3
Distribution of lexemes for the black locust (Robinia pseudoacacia) in the Sandy and sand-loamy region

lexical item	N
vergeet-mij-niet(je)	52

6.1.4
Distribution of lexemes for the forget-me-not (Myosotis arvensis) in the Dunes region

6.2 Frequency of lexemes for five plants with lowest value for internal uniformity and TTR < .2

lexical item	N	lexical item	N	lexical item	N
braambeien	1	hut bramen	1	braambeierstruik	3
braamberen	1	karrebezen	1	braambes(se)struik	4
braambezie	1	karrelbezie'nstruik	1	braambezi'n, -bezies	4
bramel	1	kattebeierboom	1	braambeier	5
kruip	1	moerbezen	1	braambeiers-, braambeier(en)hut	6
barstebeier	1	mondebeiers	1	bramers	6
bezenstruik	1	paters	1	braambees	7
braambeeshut	1	stekelbraam	1	braambeziestruik	12
braambeinen	1	struik braambezen	1	braambezelaar	16
braambessentronk	1	wilde frambozen	1	braamhut	20
braambezenbos	1	braam-, bramenhul	2	braambeiers	33
braambezenhul	1	braambees-, braambezetronk	2	braambees-, braambeze(n)struik	49
braambezietronk	1	braambessen	2	braam	58
braambreien	1	braambezebeier	2	braambezen	58
braamgewas	1	braambrei(en) struik	2	braam-, brame(n) struik	80
bramels	1	bramelhut	2	bramen	94
doorntakken	1	bramerstruik	2		
hul bramen	1	braambes	3		

6.2.1

Distribution of lexemes for the blackberry bush (Rubus fruticosus) in the Sandy and sand-loamy region

lexical item	N	lexical item	N	lexical item	N
bagweeblad	1	wever(s)kruid	1	honde-, hondstong	4
dokken	1	hondsribberen	2	wegaard(s)bladen, -blaren	5
kattestaart	1	konijneneten	2	wegbree	5
keuneblad	1	weeg-, wegeblad	2	wever(s)blaren	7
kleine wegbree	1	weewaarsbladen, -blaren	2	smalle weegbree	14
papbladen	1	wegaard(s)blad	2	weeg-, wegebree	18
ribbeplaten	1	keunoren	3	honde-, hondsrib	27
stokjes	1	smalle rib	3	rib	28
vettekerte?	1	weegbladen, -blaren, wegeb-laden, -blaren	3		
weeweeblad	1	weversblad	3		

6.2.2

Distribution of lexemes for the English plantain (Plantago lanceolate) in the Sandy and sand-loamy region

lexical item	N	lexical item	N	lexical item	N
distel	2	plakdistel	2	klevers	4
kleef	2	plakker	2	klis(se)bol	4
klitkruid	2	plakpotten	2	klis(se)kruid	4
wier	2	reit	2	klissebollen	4
bommetjes	2	smijtbollen	2	plakbollen	4
distelvinken	2	smijtdodde	2	stekkers, stekkertjes	4
doppers	2	soldate-, soldatenknop(je)	2	distels	6
dotsjes	2	stekers, stekertjes	2	kleef-, klevekruid	10
kleeftebollen	2	stekmadammetjes	2	plakkers, plakkertjes	10
klissebloem	2	sterkerbol	2	soldate(n)knoppen	10
klister	2	zoete distel	2	klissen	22
pieker	2	grote klis	4	kleefte	24
piekertjes	2	kleefbollen	4	klis	64

6.2.3
Distribution of lexemes for the lesser burdock (Arctium minus) in the Polder region

Nederlandse samenvatting

Deze dissertatie bespreekt vier casestudies over lexicale diversiteit, de hoeveelheid lexicale variatie die een concept vertoont. De hoeveelheid lexicale diversiteit kan verschillen tussen concepten. Voor een concept als DRONKEN, bijvoorbeeld, bestaan er veel meer woorden in het Nederlands, zoals *aangeschoten*, *beneveld* en *beschonken*, dan voor een concept als NUCHTER. Pilootstudies hebben aangegeven dat dat soort verschillen beïnvloed wordt door kenmerken die te maken hebben met de betekenis van de concepten zelf: de negatieve connotatie van DRONKEN, bijvoorbeeld, resulteert in een grotere mate van lexicale diversiteit (Geeraerts & Speelman 2010, Speelman & Geeraerts 2007, 2008). De pilootstudies toonden verder aan dat ook kenmerken die te maken hebben met de prototypische structuur van het lexicon, variatie in lexicale diversiteit beïnvloeden. Omdat de pilootstudies zich echter beperkten tot één semantisch veld (het menselijk lichaam) en één dialectgebied (het Limburgse dialectgebied), is de mate waarin zulke semantische kenmerken van belang zijn voor andere delen van het lexicon en voor andere variëteiten ondergedetermineerd.

De vier casestudies die aan bod komen in dit onderzoek bevestigen dat de resultaten van de pilootstudies stabiel zijn in een ander dialectgebied, omdat zowel het Limburgse als het Brabantse dialectgebied onderzocht worden, en in een diverse verzameling semantische velden. Verder leveren ze verschillende nieuwe inzichten op die nog niet behandeld werden in de pilootstudies. Het eerste deel van deze dissertatie focust voornamelijk op het bevestigen van de stabiliteit van de semantische conceptkenmerken die onderzocht werden in de pilotstudies . In de eerste casestudie (hoofdstuk 3) wordt de invloed van semantische conceptkenmerken in vijf andere semantische velden dan het menselijk lichaam, vergeleken. De semantische velden die opgenomen worden, zijn onderscheiden aan de hand van twee dimensies: de gemiddelde mate van concreetheid en de mate waarin het semantische veld lokaal, gemeenschapsgebonden of universeel is. De analyse die gepresenteerd wordt in hoofdstuk 3 bevestigt dat het effect van de conceptkenmerken uit de pilootstudies stabiel is in de andere semantische velden en in het Limburgse én Brabantse dialectgebied. De verschillen die gevonden worden tussen semantische velden betreffen enkel het feit dat bepaalde conceptkenmerken een sterkere invloed hebben in sommige velden.

De tweede casestudie (hoofdstuk 4) onderzoekt of het effect van de conceptkenmerken hetzelfde is voor de verschillende manieren waarop het geografisch gestratificeerde dialectmateriaal gekenmerkt wordt door lexicale diversiteit. Enerzijds bestaat er voor sommige concepten namelijk een grotere (of kleinere) hoeveelheid lexicale varianten. Anderzijds kunnen die varianten ook een grotere (of kleinere) mate van heterogeniteit vertonen in hun geografische distributie. Daartoe worden twee concrete onderzoeksvragen behandeld. Ten eerste bepalen we of de invloed van de conceptkenmerken significant en stabiel is als die verschillende aspecten van lexicale geografische diversiteit uit elkaar worden gehaald. Ten tweede wordt bekeken in welke mate de conceptkenmerken enkel invloed hebben *omdat* de data geografisch gestratificeerd is. De analyses tonen aan dat de onderzochte conceptkenmerken alle aspecten van lexicale diversiteit beïnvloeden, hoewel de effecten van de verschillende kenmerken niet even sterk zijn voor elk aspect van lexicale diversiteit. Verder kunnen we besluiten dat de geografische verdeling van het dialectmateriaal niet de enige reden is waarom de conceptkenmerken lexicale diversiteit beïnvloeden.

In het tweede deel van deze verhandeling wordt het feit dat semantische kenmerken ook lectale variatie kunnen vertonen, in beschouwing genomen. Zo is het, bijvoorbeeld, waarschijnlijk dat een concept voor een bepaalde spreker beter bekend is, omdat hij/zij er frequenter mee in aanraking

komt. De derde casestudie (hoofdstuk 5) onderzoekt het gebruik van leenwoorden in dialectgebieden vanuit onomasiologisch perspectief. Het hoofdstuk focust, meer bepaald, op de interactie tussen de geografische locatie van een taalgebruiker en semantiek als verklaring van de frequentie van leenwoorden uit drie brontalen (Frans, Latijn en Duits) in vier semantische velden in Limburg en Brabant. De resultaten tonen aan dat geografie een belangrijke rol speelt, omdat er grenseffecten optreden. Duitse leenwoorden komen, bijvoorbeeld, frequenter voor nabij de grens met Duitsland in het Limburgse dialectgebied. Verder blijkt ook de interactie met semantiek een cruciale rol te spelen: het gebruik van Franse leenwoorden is, bijvoorbeeld, in grote mate afhankelijk van het semantische veld waartoe het concept behoort en van de locatie van de dialectspreker. De patronen die gevonden worden, kunnen enkel verklaard worden aan de hand van de socio-culturele achtergrond van de dialectsprekers.

In de laatste casestudie (hoofdstuk 6) wordt rechtstreeks onderzocht wat het effect is van de dagelijkse omgeving van een taalgebruiker op lexicale diversiteit. De casestudie zoomt in meer detail in op het semantische veld van planten in de Brabantse, Limburgse en Vlaamse dialecten in België. Het presenteert een analyse van de relatie tussen lexicale diversiteit in een bepaalde regio per plant en de frequentie waarmee die plant voorkomt in de omgeving van de dialectsprekers uit die regio. Cruciaal daarbij is dat plantfrequentie afhankelijk is van de locatie van de spreker: in de kempen komen, bijvoorbeeld, meer heidevelden voor dan in andere gebieden in België. De analyse toont aan dat zulke extra-linguïstische factoren correleren met lexicale diversiteit: hoe frequenter de plant, hoe minder lexicale diversiteit er gevonden wordt. Andere factoren, zoals de frequentie waarmee een plant cultureel relevant is, bijvoorbeeld omdat hij voorkomt in volksgeloof, moeten echter ook een rol spelen.

Over het algemeen geeft deze studie dus een systematisch en gevarieerd beeld van het dialectlexicon van het Nederlands. Naast het feit dat de dialecten geografisch gestratificeerd zijn, wordt duidelijk aangetoond dat lexicale verschillen tussen verschillende locaties kleiner (of groter) zijn, afhankelijk van de betekenis van de referent waarnaar verwezen wordt.